Expert Systems With Applications 231 (2023) 120676

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Symbolic regression with feature selection of dye biosorption from an aqueous solution using pumpkin seed husk using evolutionary computation-based automatic programming methods

Sibel Arslan[a,*], Nurşah Kütük[b]

[a] Department of Software Engineering, Sivas Cumhuriyet University, Sivas, 58140, Turkey
[b] Department of Chemical Engineering, Sivas Cumhuriyet University, Sivas, 58140, Turkey

ARTICLE INFO

Keywords:
Pumpkin seed husk
Biosorption
Titan yellow
System modeling
Genetic programming
Artificial bee colony programming

ABSTRACT

Industrial waste pollution is a serious and systematic problem that harms the environment and people. The development of cheap, simple, and efficient techniques to solve this problem is important for sustainability. In this study, both experimental and evolutionary computation (EC)-based automatic programming (AP) methods were used to investigate the biosorption process for water treatment. In the experiments, titan yellow (TY), an anionic dye, was biosorbed from an aqueous solution containing pumpkin seed husk (PSH). The structure of PSH was examined using a Fourier transform infrared spectroscopy (FTIR) and a scanning electron microscope (SEM). The result of the experimental studies was that TY biosorption of PSH reached a biosorption efficiency of 95% after 120 min of contact time. The maximum biosorption capacity (q_{max}) was calculated to be 181.8 mg/g. It was found that the biosorption of TY better followed the Dubinin–Radushkevich isotherm ($R^2 = 0.98$) and pseudo second-order reaction kinetics ($R^2 = 0.99$). Based on the thermodynamic data, the biosorption process was exothermic and spontaneous. After the experiments, the process was modeled using pH, biosorbent concentration, initial dye concentration, contact time, and temperature as inputs and biosorption efficiency (%) as output for the methods. Moreover, the success of these AP methods was compared with a newly proposed evolutionary method. The simulation results indicate that AP methods generate best models ($R^2_{\text{train}} = 0.99$ and $R^2_{\text{test}} = 0.97$). At the same time, the most important parameter of the process in the feature analysis is contact time. This study shows that EC-based AP methods can effectively model applications such as the biosorption process.

1. Introduction

Environmental pollution has become dangerous worldwide due to rapidly increasing urbanization and industrialization (Bahramian, Dereli, Zhao, Giberti, & Casey, 2022; Saravanan et al., 2019). Dye wastes from various sectors such as textiles, plastics, and paints, especially polluting water bodies, cause severe damage to the aquatic ecosystem, environment, and human health due to their toxic, mutagenic, and carcinogenic effects (Chaurasia, Jasuja, & Kumar, 2022; Tabaraki & Sadeghinejad, 2018). For this reason, researchers from around the world have been working in recent years on ways to remove organic pollution from water (Vidya, Manjunatha, Sudeep, Ashoka, & Raj, 2020). Organic dyes that cause this pollution are wastes with complex chemical, aromatic, and stable structures that are difficult to break down in nature (Shi, Li, Wang, Wang, & Cao, 2020; Tabaraki & Sadeghinejad, 2018). Even in low concentrations, the release of dye waste into nature causes serious ecological problems (Rigueto et al., 2021). Azo dyes are used in various fields such as paints, paper, plastics, textiles and leather (Bameri, Saffari, Baniyaghoob, & Ekrami-Kakhki, 2022; Ibrahim, Allah, & Muneer, 2021; Vidya et al., 2020). Titan yellow (TY), an anionic azo dye, is an organic dye molecule with sulfonic groups (Ibrahim et al., 2021; Shi et al., 2020). While TY causes serious environmental damage in wastewater, it can damage the digestive system, eyes and skin upon contact (Hiremath, Mal, Prabha, & Vidya, 2018; Shi et al., 2020). There are various methods for the removal of dyes such as filtration, osmosis, coagulation, membrane processes, and oxidation. However, these methods are expensive when used continuously. Adsorption technique, on the other hand, is a cost-effective technique with high efficiency in dye removal (Bameri et al., 2022; Tabaraki & Sadeghinejad, 2018). Another alternative method is biosorption, which can use various biomasses such as bacteria, fungi and plants (Rigueto et al., 2021; Tabaraki & Sadeghinejad, 2018).

* Corresponding author.
E-mail addresses: sibelarslan@cumhuriyet.edu.tr (S. Arslan), nkutuk@cumhuriyet.edu.tr (N. Kütük).

https://doi.org/10.1016/j.eswa.2023.120676
Received 1 February 2023; Received in revised form 31 May 2023; Accepted 31 May 2023
Available online 7 June 2023

There have been numerous studies on modeling the adsorption/biosorption process and the parameters affecting this process using different methods. Tabaraki et al. synthesized arginine-modified magnetic Fe_3O_4/chitosan nanoparticles for TY adsorption. As a result of the adsorption process, 52% dye removal and 374 mg/g adsorption capacity were achieved. Moreover, the response surface method (RSM) was used with the Box–Behnken model (Tabaraki & Sadeghinejad, 2018). Moreira et al. dried *Chlorella pyrenoidosa* microalgae were modeled using RSM and artificial neural network (ANN) to optimize experimental conditions for biosorption of copper. Modeling results showed that RSM had better accuracy than ANN (Moreira, Lebron, & de Souza Santos, 2020). In a study where Hypnea musciformis powder was used as a biosorbent, the biosorption of TY dye was optimized using the central composite design program. Optimization values for the parameters were proposed to achieve maximum biosorption by the model (Raju & Sunil, 2018). In another study, in which a functionalized copper oxide-zinc oxide nanocomposite was synthesized as an adsorbent, the adsorption process was modeled by genetic programming (GP) (Mahmoodi, Chamani, & Kariminia, 2016).

Although previous studies have mostly used conventional methods to model biosorption processes, selecting relevant features for the process can be difficult even in such a specialized research area as evolutionary computation (EC). The increasing complexity of engineering problems and the inability of mathematical methods to find optimal solutions have significantly increased the use of EC-based algorithms. Whale optimization algorithm (Mirjalili & Lewis, 2016), firefly algorithm (FA) (Yang, 2008), cuckoo search algorithm (CSA) (Yang & Deb, 2009), bee colony optimization (BCO) (Teodorovic, Lucic, Markovic, & Dell'Orco, 2006), and dragonfly algorithm (DFA) (Mirjalili, 2016) can be cited as examples of algorithms based on EC, and these can be used by hybridizing them with different methods. Examples of such applications include precipitation index prediction using the ANN-FA hybrid model (Mohammadi, 2023), solar radiation estimation using the CSA based hybrid model (Moazenzadeh, Mohammadi, Duan, & Delghandi, 2022), and daily dew point temperature prediction using the BCO, DFA, and Adaptive Neural-Fuzzy Inference Systems (ANFIS) hybrid model (Mehdizadeh, Mohammadi, & Ahmadi, 2022).

By solving the symbolic regression (SR) problem, the mathematical relationship between the inputs and outputs of a system is determined. Recent developments in artificial intelligence have led to renewed interest in SR, and several methods have been developed. One of these methods is the reinforcement learning approach, in which the recurrent neural network is trained with a risk-seeking policy gradient (Petersen et al., 2019). The method called deep symbolic regression (DSR) has outperformed many baseline methods and proved that deep learning can be adapted to SR problems. Another method is iterative variable selection assisted SISSO (VS-SISSO) method, which aims to select meaningful features by adapting to high-dimensional datasets (Guo, Hu, Han, & Ouyang, 2022). The purpose of the method is to keep the dimensions of the models constant and try to choose the model that best describes the data by selecting features from subsets of features each time. In this method, the size of the selected subspace changes depending on the number of dimensions. This method is similar to the idea of GP. However, unlike GP, VS-SISSO evaluates features as individuals rather than as expressions extracted from solution trees. Another GP-like method is DoME (Rivero, Fernandez-Blanco, & Pazos, 2022). It uses solutions in tree structures as in GP and generates the model by computing the error value for each node and finding the best constant for the node. SR-Forest, on the other hand, combines the advantages of GP and decision trees (Zhang, Zhou, Chen, Xue, & Zhang, 2023). With these advantages, it also creates a SR-based ensemble model using the mutation operator to deal with high-dimensional datasets. Difference-based firefly programming (DFP) is also a EC-based automatic programming (AP) method, similar to GP (Aliwi, Demirci, & Aslan, 2023). DFP, which produces parse tree solutions like GP, has improved standard firefly programming with simplification and substitution operators.

The biosorption process can be modeled as a SR problem because it involves the analysis of the data generated in the experiments, which requires a lot of time and effort. Moreover, feature importance (FI) analysis can be performed to explain the output of the biosorption models. The contribution of each input in the model to the model output can be predicted by FI (Rengasamy, Rothwell, & Figueredo, 2021). In the literature, many FI methods are adapted to various machine learning algorithms (Oh, 2022; Wei, Zhao, Feng, He, & Yu, 2020). However, some are model independent and some are model specific (Rengasamy et al., 2022).

The most commonly used EC based-AP methods for solving SR problems and determining FI are GP and artificial bee colony programming (ABCP) (Moghaddam, Al-Sahaf, Xue, Hollitt, & Zhang, 2021; Nekoei, Moghaddas, Golafshani, & Gandomi, 2021; Yamashita, Fogliatto, Anzanello, & Tortorella, 2022; Zojaji, Ebadzadeh, & Nasiri, 2022). Therefore, in our study we modeled the process using successful versions of these methods, multi-gene GP (MGGP) (Kütük & Arslan, 2022) and multi-hive ABCP (MHABCP) (Arslan & Ozturk, 2019b), and attempted to select the most important parameter for the process. The stages and contributions of our study are as follows:

- The biosorption process by selecting TY, which is used in industry as a model dye compound. We used pumpkin seed husk (PSH) as a biosorbent because it is cheap, abundant, and has a fibrous structure (Hameed & El-Khaiary, 2008).
- The structure of PSH before and after biosorption (FTIR and SEM) was examined.
- The parameters (pH, biosorbent concentration, initial dye concentration, contact time, and temperature) affecting the biosorption process were investigated. Isotherm, reaction kinetic and thermodynamic calculations were performed with the data obtained from these parameters.
- The process was modeled using EC based AP methods and these methods were compared with a recently proposed evolutionary algorithm.
- The models were analyzed and the effects of the parameters on the process were investigated.
- To the best of our knowledge, this study was the first to model the biosorption process with MHABCP.

As outlined above, there is a research gap in the use of AP methods in chemical engineering, especially in process modeling. This pioneering and detailed study will fill this gap by proposing the most accurate models to correctly select the relevant parameters of the biosorption process. This will make an important contribution to researchers interested in chemical engineering and EC.

We organize the remainder of this study as follows. Section 2 describes the material and method of the biosorption process. This section also includes the definitions of the AP methods used in the modeling. Section 3 presents in detail the analysis of the data obtained by the experiments, the performance evaluation criteria, and the parameters of the methods. The analysis of the biosorption process, the simulation results of the models generated by the methods, the feature analysis, and additional study comparing AP methods are described in Section 4. Research questions about this study were answered in Section 5. Finally, we conclude this study in Section 6.

2. Materials and methods

2.1. Materials

The biosorbent used in the study, PSH, was supplied from local markets. TY (%99 purity) was obtained from Eastman Organic Chemicals. Hydrochloric acid (HCl, %33) and sodium hydroxide (NaOH, pellet) were obtained from Kimetsan (Ankara, Turkey). The physicochemical properties of TY are given in Table 1.

Table 1
Properties of dye.

Dye	Titan yellow
Chemical formula	$C_{28}H_{19}$ $N_5Na_2O_6$ S_4
Molecular weight (g/mol)	695.720 g/mol
Ionic structure	Anionic
Solubility	High (for water)
Color	Yellow
λ_{max}	405 nm

2.2. Biosorption procedure

To increase the surface area, the size of the PSH was reduced by 0.05–0.1 mm using a grinding machine (Sinbo, coffee grinder, SCM-2934). First, an aqueous stock solution (1000 mg/L) of TY was prepared. The dilution was continued according to the concentration to be used later. The experiments were performed intermittently in a 250 mL glass flask and 100 mL dye volume. Samples taken at specified time intervals were filtered. Then, the absorbance values were determined at 405 nm by UV/vis spectroscopy. In the biosorption studies, dye removal was optimized by the parameters of pH (2–8), initial dye concentration (10–300 mg/L), biosorbent concentration (0.5–10 g/L), contact time (0–120 min) and temperature (20–50 °C).

$$\% \text{ Biosorption efficiency } = (C_o - C)/C_o \times 100 \tag{1}$$

$$q_e = ((C_o - C_e).V)/m \tag{2}$$

$$q_t = (((C_o - C_t) \cdot V))/m \tag{3}$$

In Eq. (1), C_0 is the initial concentration of the dye (mg/L) and C is the concentration (mg/L) at time t. In Eqs. (2) and (3), q_e is the biosorption capacity (mg/g) at equilibrium, q_t is the biosorption capacity (mg/g) at t = t, C_e is the final dye concentration (mg/L), V is the solution volume (mL), and m is the biosorbent amount (g). The absorbance values of the TY dye solution were determined by UV spectroscopy (Schimadzu, 1601).

2.3. Characterization

2.3.1. Characterization of biosorbent

Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR, Bruker, Tensor II) spectroscopy was used to analyze the chemical structure of the PSH and TY dyes while scanning electron microscopy was analyzed the morphological structure (SEM, Tescan Mira 3 XMU). FTIR spectroscopy was examined milled materials in the 4000–400 cm^{-1} wavenumber region.

2.3.2. The point of zero charges (pH_{pzc})

The method of examining the initial and equilibrium pH values was used to determine the zero charge (pH_{pzc}) potential of PSH. It was studied in the amount of 100 mg biosorbent in the pH range of 2–10, at room temperature and for 120 min with stirring. pH levels were adjusted with NaOH and HCl solutions. The difference (ΔpH) was determined by making initial and final pH measurements. Initial pH value and (ΔpH) value were plotted and zero charge point (pH_{pzc}) was determined with the help of graph.

2.3.3. Biosorption isotherm models used

Isotherms help to understand the maximum biosorption capacity of biosorption and the type of biosorption. Langmuir isotherm (Eq. (4)) indicates that the surface of the biosorbent is homogeneous and the presence of monolayer biosorption. R_L is a coefficient indicating the suitability of the Langmuir isotherm Eq. (5) Freundlich isotherm (Eq. (6)), on the other hand, refers to multilayer biosorption and the heterogeneous structure of the biosorbent surface (Kütük & Arslan, 2022; Rattanapan, Srikram, & Kongsune, 2017). Dubinin-Radushkevich (D-R) (Eq. (7)) isotherm can be evaluated to get an idea of how the biosorbent and dye interact (Isik, Ugraskan, & Cankurtaran, 2022; Rattanapan et al., 2017). In Eq. (4), q_{max} (mg/g) is the maximum biosorption capacity, K_L (L/mg) is the Langmuir constant, and C_e is the dye concentration at equilibrium. In Eq. (6), K_f (L/g) is the Freundlich constant, and n is the constant expressing the adsorption capacity density. A value of n between 0 and 1 indicates that the isotherm is suitable. Eq. (7), k_D (mol^2/J^2) is the constant of the D-R isotherm, which expresses the free biosorption energy per mole of the dye. Polanyi potential, ε is calculated according to Eq. (8). In Eq. (9), the E value represents the average biosorption energy. If the E value is 8–18 kJ mol^{-1}, it can be said that there is ion exchange, if it is less than 8 kJ mol^{-1}, there is physical adsorption, and if it is between 20–40 kJ mol^{-1}, there is chemical adsorption (Isik et al., 2022).

$$\frac{1}{q_e} = \frac{1}{q_{max}} + \frac{1}{K_L q_{max}} x \frac{1}{C_e} \tag{4}$$

$$R_L = \frac{1}{1 + K_L \cdot C_0} \tag{5}$$

$$\ln q_e = \ln K_f + \frac{1}{n} x \ln C_e \tag{6}$$

$$Lnq_e = Lnq_{\max} - k_D \cdot \varepsilon^2 \tag{7}$$

$$\varepsilon = RT \,\text{Ln}\left(1 + \frac{1}{C_e}\right) \tag{8}$$

$$E = \frac{1}{\sqrt{2K_D}} \tag{9}$$

2.3.4. Kinetic models used

Biosorption kinetics studies were investigated with 4 different models. In Eqs. (10), (11), (12) and (13), pseudo-first-order kinetic model (PFO), pseudo-second-order kinetic model (PSO), intraparticle diffusion model and Elovich model are defined, respectively. q_t (mg/g) is the biosorption capacity at time t. k_1 (1/min), k_2 (g/mg min) and k_i (mg/g min^2) which are the rate constants for PFO, PSO and the diffusion model for the particle, respectively. Elovich model is known as α (mg/g min), and the desorption rate constant, β (g/mg), is related to the activation energy for chemical adsorption (Bameri et al., 2022; Netzahuatl-Muñoz, Guillén-Jiménez, Chávez-Gómez, Villegas-Garrido, & Cristiani-Urbina, 2012; Rigueto et al., 2021).

$$\log\left(q_e - q_t\right) = \log q_e - \frac{k_1}{2.303}\text{t} \tag{10}$$

$$\frac{t}{q_t} = \frac{1}{k_2 \cdot q_e 2} + \frac{1}{q_e} t \tag{11}$$

$$q_t = k_{id} \cdot t^{1/2} + \text{C} \tag{12}$$

$$q_t = \frac{1}{\beta}\ln(\alpha \cdot \beta) + \frac{1}{\beta}\ln t \tag{13}$$

2.3.5. Thermodynamic characteristics

Thermodynamic data is used to determine the mechanism of biosorption and its potential to occur spontaneously (Isik et al., 2022). Thermodynamic parameters enthalpy energy (ΔH, kJ/mol), entropy change (ΔS, kJ/mol K) and free energy change (ΔG, kJ/mol) data were calculated according to Eqs. (14)–(16).

$$K_c = C_a/C_e \tag{14}$$

$$\text{In } K_C = \frac{\Delta \text{S}}{\text{R}} - \frac{\Delta H}{R} \cdot \frac{1}{T} \tag{15}$$

$$\Delta G = \Delta H - T\Delta S \tag{16}$$

K_c is the equilibrium constant, C_a is the amount of dye retained in unit mass of biosorbent (mg/g), C_e is the dye concentration remaining in solution (mg/L). Ideal gas constant R is taken as 8.314 J/mol K.

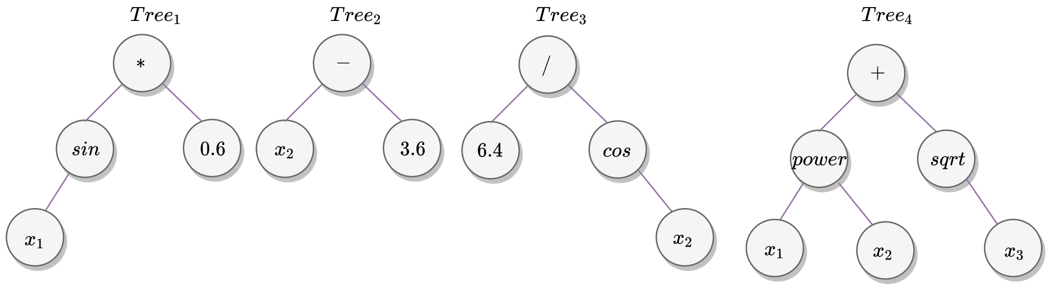


Fig. 1. Example of multi tree solution in MGGP and MHABCP.

2.4. AP methods

AP, one of the subfields of machine learning, generates nonlinear mathematical models that represent systems (Arslan & Koca, 2023). It also attempts to predict the relationship between the input and output parameters of systems with the least error. In this way, information between the system's inputs and outputs can be discovered. Many EC-based AP methods have been attracting more and more attention worldwide. Two pioneers of these methods are GP (Koza, 1994) and ABCP (Karaboga, Ozturk, Karaboga, & Gorkemli, 2012).

The commonly used versions of GP and ABCP are MGGP and MHABCP that produce multi-tree solutions that have been successfully applied to many complex problems such as SR (Adeyi et al., 2022; Boudouaoui, Habbi, Ozturk, & Karaboga, 2020; Datta, Dev, & Eden, 2019; Ge, Yusa, & Fan, 2021), feature selection (Arslan & Ozturk, 2019b; Öztürk, Tarım, & Arslan, 2020), prediction (Kazemi & Barati, 2022; May Tzuc et al., 2019; Pawanr, Garg, & Routroy, 2022; Sattar, Majid, Kausar, Bilal, & Kashif, 2022), process modeling (Punugupati, Kandi, Bose, & Rao, 2017), classification (Arslan & Ozturk, 2018, 2019a; Pedrino, Yamada, Lunardi, & de Melo Vieira, 2019) and forecasting (Hadi & Tombul, 2018).

The main advantage of MGGP and MHABCP is that the system can automatically generate the model form with coefficients and variables, since each tree in the solution has a nonlinear structure. The smallest unit of the trees, the nodes, is selected from either the terminal set or the function set. While the terminal set consists of variables and constants like x, y, z; the function set consists of arithmetic operators like $+$, $-$, $*$, $/$ and mathematical functions like sin, cos, e^x, $sqrt$ and $power$. An example of a solution for these methods with multi-tree structures is shown in Fig. 1. The mathematical formula for the solution shown in the figure is given in Eq. (17). Flowcharts of the methods are shown in Fig. 2.

$$\begin{aligned} \mathrm{y} &= \mathrm{w}_0 + \mathrm{w}_1 \times \text{ Tree }_1 + \mathrm{w}_2 \times \text{ Tree }_2 + \mathrm{w}_3 \times \text{ Tree }_3 + \mathrm{w}_4 \times \text{ Tree }_4 \\ &= w_0 + w_1 \times 0.6 \sin\left(x_1\right) + w_2 \times (x_2 - 3.6) + w_3 \times \frac{6.4}{\cos\left(x_2\right)} \\ &\quad + w_4 \times \left(x_1^{x_2} + \sqrt{x_3}\right) \end{aligned} \tag{17}$$

In both algorithms, the trees are generated using the ramped half-and-half method. In this method, half of the solutions are generated with the full method and the other half with the grow method. In the full method, the nodes are selected from the function set up to the desired maximum depth and the remaining nodes are selected from the terminal set. In the grow method, the maximum tree depth is achieved by randomly selecting from both the function set and the terminal set. Since ramped half and half contains solutions generated by both methods, diversity in the population/colony is guaranteed. Koza suggested that the generated solutions should be different from each other (Koza, 1994). In our study, we ensured that the solutions in the initial population/colony were not repeated by either method.

In the initial phase of MHABCP, the solutions in the colony are randomly generated from hives between 1 and the maximum number of hives, $H_{\max}$ (Arslan & Ozturk, 2019b). The quality of each solution is determined by considering the fitness function specific to the problem. In the MHABCP algorithm, there are bees with three different tasks: employed bees, onlooker bees, and scout bees. Two basic operators are used to try to improve the bees/solutions. The first is the information sharing mechanism, which is also proposed in ABCP (Fig. 3). The other operator is the MGGP-inspired hive exchange mechanism (Fig. 4). Since the same operators with different names are also found in MGGP, these operators are mentioned after the explanation of MGGP.

When the employed bees leave the nest, they have a specific food source in their memory and when they return to the nest, they share information with the onlooker bees about the food sources. In this phase, each employed bee tries to improve itself through the mechanisms. After this phase is completed, the probability of selecting solutions is calculated based on their fitness values. The onlooker bees evaluate the information shared by the employed bees and decide which food source they will go to based on the amount of nectar in the food source. Each onlooker bees tries to improve itself through the two mechanisms mentioned above.

With counters initially set to 0, each solution is checked to determine how many unsuccessful improvements have been made. Whether the solution is abandoned or not is determined by the $limit$ parameter. The counter is incremented by 1 for each unsuccessful improvement attempt. If there is a solution whose counter value is greater than $limit$ after the onlooker bee phase, this solution will be abandoned in the scout bee phase. Unlike other types of bees, these bees search for new food sources without sharing information. After they find the food sources, the counter value of the solutions for these sources is set to 0 and the scout bees continue their work as employed bees.

The steps of the MGGP algorithm are similar to those of GP, but there are differences in the operation of the optimization operators due to the tree structure of the solutions. First, initial solutions are generated and the best solution(s) are stored using elitism. This operator ensures that a certain number of the best solutions are passed on to the next generation. Thus, the best individuals in the population are retained over generations. After elitism, attempts are made to improve the solutions with three evolutionary operators: selection, crossover, and mutation. Each of these operators is applied to every solution in the population. The selection operator is used to select parents that generate new solutions based on the fitness values of the solutions in the population. This operator is used to evaluate the fitness of each individual in the population for the problem. Individuals with good fitness are more likely to be recovered in the next generation. It is used with methods such as the selection operator, the roulette wheel, or the tournament method. The crossover operator (Fig. 3) randomly modifies selected subtrees of parent solutions to generate child solutions. The high-level crossover operator (Fig. 4) proposed by MGGP exchanges the trees (genes) that represent the solution. In the mutation operator, a selected node or subtree is changed depending on the mutation probability.

As mentioned earlier, two important operators in MGGP are adapted to MHABCP. The first is the modifier crossover operator, which is called the information sharing mechanism in MHABCP. The other operator is

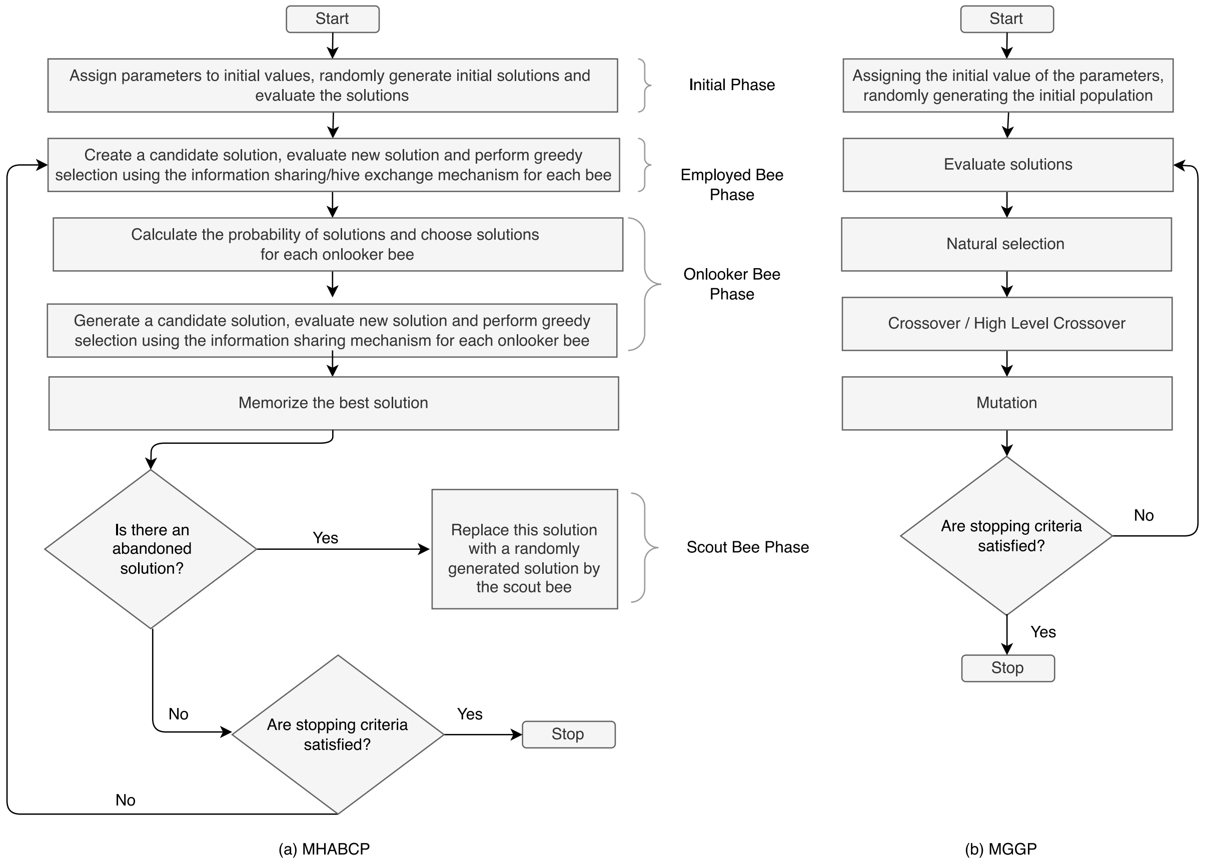

Start
Assign parameters to initial values, randomly generate initial solutions and evaluate the solutions
Initial Phase
Create a candidate solution, evaluate new solution and perform greedy selection using the information sharing/hive exchange mechanism for each bee
Employed Bee Phase
Calculate the probability of solutions and choose solutions for each onlooker bee
Onlooker Bee Phase
Generate a candidate solution, evaluate new solution and perform greedy selection using the information sharing mechanism for each onlooker bee
Memorize the best solution
Is there an abandoned solution?
Yes
Replace this solution with a randomly generated solution by the scout bee
Scout Bee Phase
No
Are stopping criteria satisfied?
Yes
Stop
No
(a) MHABCP
Start
Assigning the initial value of the parameters, randomly generating the initial population
Evaluate solutions
Natural selection
Crossover / High Level Crossover
Mutation
Are stopping criteria satisfied?
No
Yes
Stop
(b) MGGP

Fig. 2. Flowcharts of AP methods.

Table 2
The comparison of the methods used.

Criteria	MGGP	MHABCP
Initialization	Ramped half and half	Ramped half and half
Operator	High-level crossover, crossover, mutation	Hive exchange mechanism, information sharing mechanism
Solution trees	Multi gen	Multi hive
Inspired algorithm	GP	ABCP

the high-level crossover operator, which is called the hive exchange mechanism. The representation of the two operators can be found in Figs. 3 and 4.

Both algorithms improve the solutions with their own improvement mechanisms. These processes run until criteria such as the number of iterations and the number of evaluations are satisfied. The comparison of the methods can be found in Table 2. It should be noted that although both algorithms use the same technique for generating output solutions, the flowcharts and operators for improving the solutions in this table differ.

3. Experimental design

In this section, we explained the generated dataset and presented information about the fitness functions and parameters.

We expect to answer the following research questions with experiments:

- Can PSH be used as an efficient biosorbent for the removal of TY dye from wastewater?
- Under which experimental conditions can the highest biosorption efficiency be achieved when the process is optimized?
- Which isotherm and kinetic models is the biosorption reaction compatible with?
- Can EC -based AP methods be used to model water treatment processes such as biosorption?
- What is the observed error value when the biosorption process is mathematically modeled using different methods?
- In what order are the features that contribute to the process important?

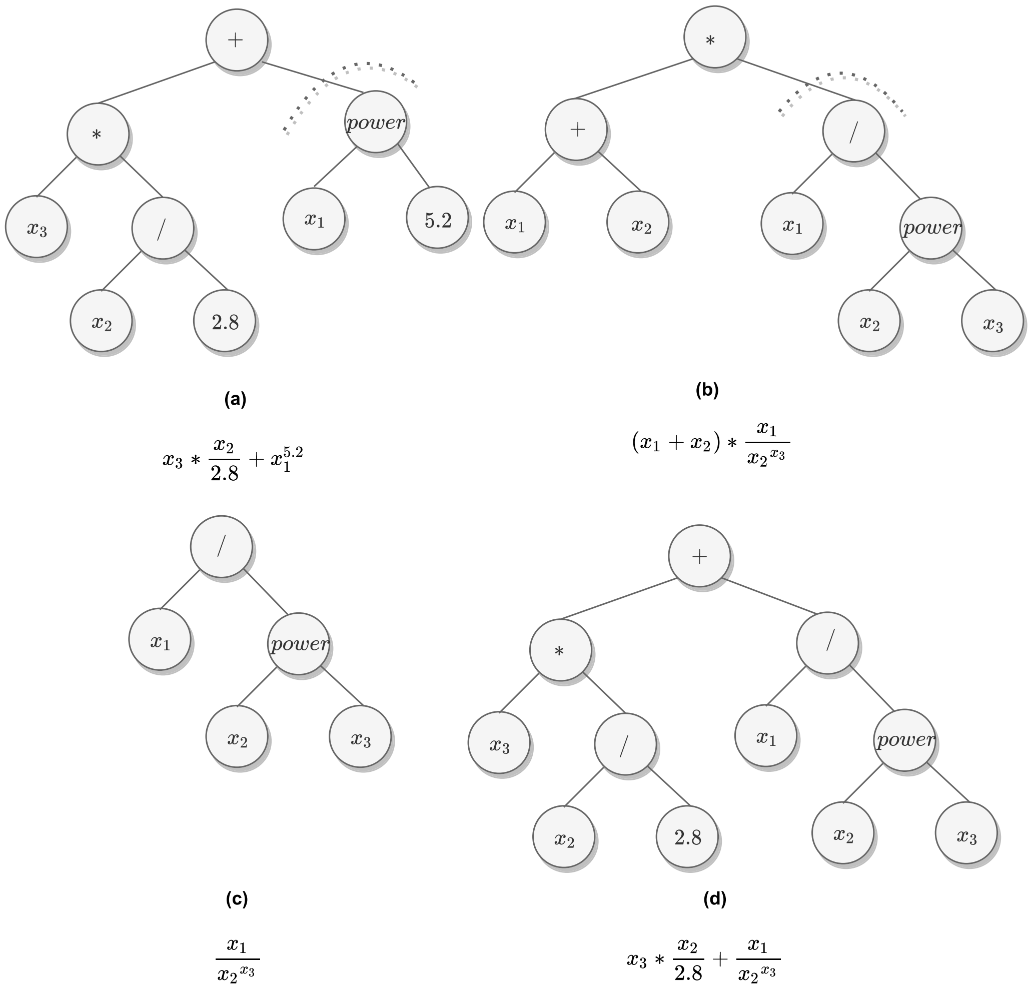
3.1. Data and analysis

The experimental dataset was generated under laboratory conditions to improve predictive models. It was partitioned based on the data partitioning generally recommended in machine learning (70% training and 30% test samples) (Alizadeh, Shahheydari, Kavianpour, Shamloo, & Barati, 2017).

The total number of samples in the dataset is 84, so 59 randomly selected samples (about 70%) were used for training and the remaining 25 samples (about 30%) were used for testing. Each data sample has 6 parameters: (x_1) pH, (x_2) biosorbent concentration (g/L), (x_3) initial dye concentration (mg/L), (x_4) contact time (min), (x_5) temperature (°C), and (y) biosorption efficiency (%). The biosorption is the actual output (y) and should be predicted from other parameters. It can be noted that all features except output are represented as input parameters for the methods to predict the actual output of the samples. This representation can be seen in Fig. 5.

3.2. Performance evaluation criteria

Methods were evaluated using the sum squared error (SSE), root mean square error (RMSE), and mean absolute error (MAE) fitness



(a)

$x_3 * \frac{x_2}{2.8} + x_1^{5.2}$

(b)

$(x_1 + x_2) * \frac{x_1}{{x_2}^{x_3}}$

(c)

$\frac{x_1}{{x_2}^{x_3}}$

(d)

$x_3 * \frac{x_2}{2.8} + \frac{x_1}{{x_2}^{x_3}}$

Fig. 3. Information sharing mechanism of MHABCP/Crossover operator MGGP ((a): current solution, (b): neighborhood solution, (c): sub tree selected from neighborhood solution, (d): candidate solution).

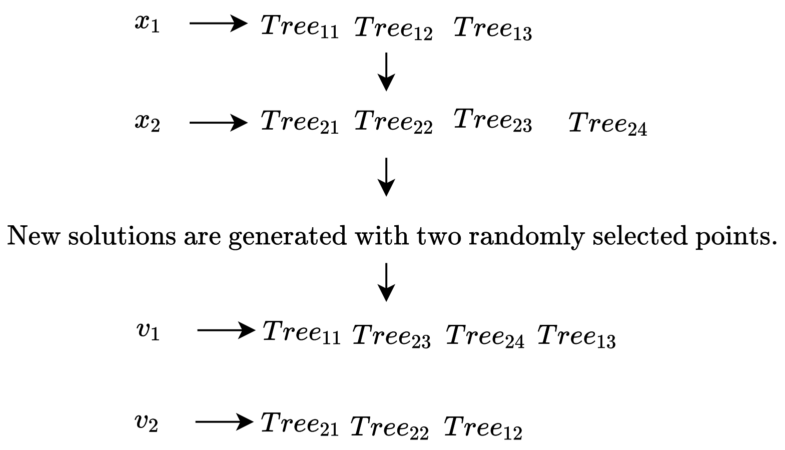

$x_1 \longrightarrow Tree_{11}\ Tree_{12}\ Tree_{13}$
$x_2 \longrightarrow Tree_{21}\ Tree_{22}\ Tree_{23}\ Tree_{24}$
New solutions are generated with two randomly selected points.
$v_1 \longrightarrow Tree_{11}\ Tree_{23}\ Tree_{24}\ Tree_{13}$
$v_2 \longrightarrow Tree_{21}\ Tree_{22}\ Tree_{12}$

Fig. 4. High-level crossover of MGGP/hive exchange mechanism of MHABCP.

functions. SSE, RMSE, and MAE are given in Eqs. (18), (19), and (20), respectively.

$$SSE = \sum_{i=1}^{N} \left(f\left(x_i\right) - Y_i\right)^2 \tag{18}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(f\left(x_i\right) - Y_i\right)^2} \tag{19}$$

$$MAE = \frac{\sum_{i=1}^{N} \left|f\left(x_i\right) - Y_i\right|}{N} \tag{20}$$

where $f\left(x_i\right)$ is predicted output of models proposed by the methods, Y_i is output of generated by real experiments and N is the number of samples. Moreover, the performance of the models was statistically evaluated by the coefficient of determination (goodness of fit, R^2) using Eq. (21). The closer this value is to 1, the better the actual data fit the

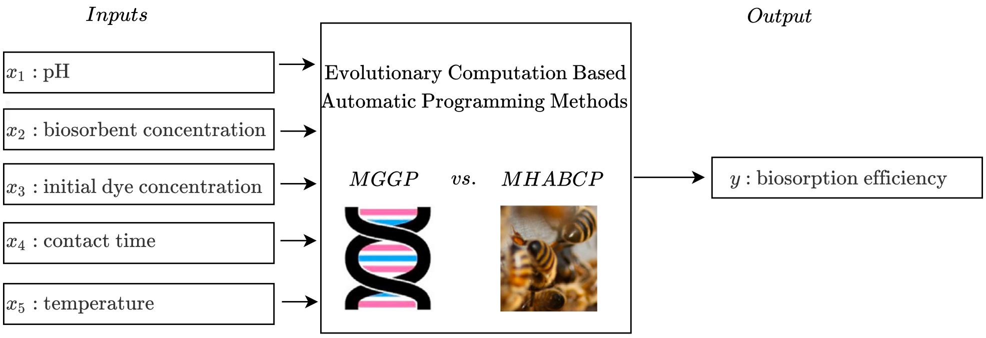

Inputs
x_1 : pH
x_2 : biosorbent concentration
x_3 : initial dye concentration
x_4 : contact time
x_5 : temperature
Evolutionary Computation Based Automatic Programming Methods
MGGP vs. MHABCP
Output
y : biosorption efficiency

Fig. 5. Data analysis of the proposed methods.

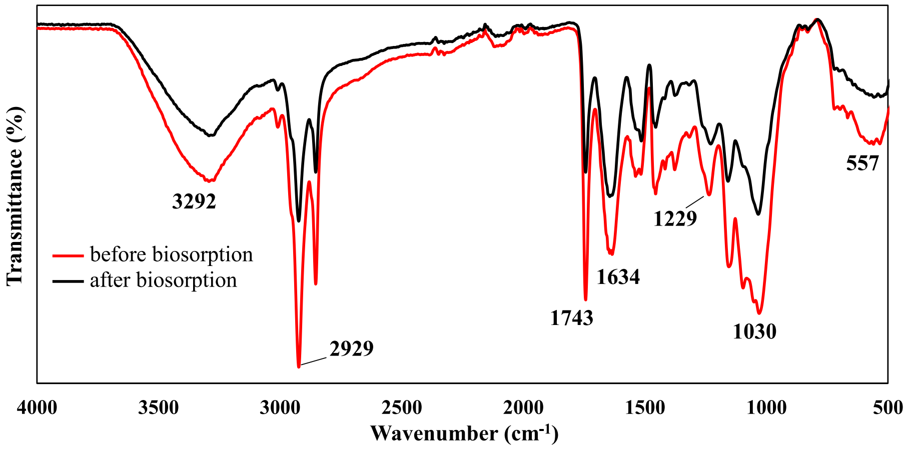


Fig. 6. FTIR spectrum of PSH biosorption with TY.

predicted data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(f\left(x_i\right) - Y_i\right)^2}{\sum_{i=1}^{N} \left(Y_i - Y\right)^2} \tag{21}$$

The complexity of trees of the models C is also proportional to the depth of solution d and the total number of nodes in the corresponding depth of solution n, and is calculated as in Eq. (22).

$$C = \sum_{k=1}^{d} n * k \tag{22}$$

3.3. Parameters of AP methods

Since parameter selection also affects the success of the algorithm, we performed hyperparameter optimization using the Grid Search (GS) algorithm for the parameters determined in both algorithms. GS tries all possible combinations in a given hyperparameter space to select the best performing parameters. For population/colony size (100, 200, 250, 500), number of iterations (100, 200, 250, 500), tournament size (5, 10, 20), information sharing mechanism/high level cross rate (0.1, 0.2, 0.5), and maximum number of genes/hives (2, 3, 5, 10). In total, the parameters were adjusted among $4 * 4 * 3 * 3 * 4 = 516$ different combinations and other parameters determined by expert opinions are shown in Table 3.

For a fair comparison, the same parameter values are used in MGGP and MHABCP. To improve the solutions, high-level crossover, crossover, mutation, and direct reproduction were used in MGGP, while information sharing mechanism and hive exchange mechanism operators were used in MHABCP. Terminals express the functions *square* (squared), *cube* (cubed), *power* (exponent), *neg* (negative), and *sqrt* (square root). They are also functions representing the product of the three terminals *mult*3 and their sum *add*3. Other functions in Table 3 give the trigonometric function equivalent of the terminal.

Table 3
Parameters.

Parameters	MGGP	MHABCP
Population size	250	–
Colony size	–	250
Generation	500	500
Maximum tree depth	5	5
Crossover rate	0.84	–
Mutation rate	0.14	–
Direct reproduction Rate	0.02	–
Tournament size	20	–
Limit	–	50
Maximum number of gen/hive	10	10
Information sharing mechanism rate	–	0.8
Hive exchange mechanism rate	–	0.2
High level crossover	0.2	–
Functions	+, −, ∗,/ (protected), *square*, *cube*, *sin*, *cos*, *add*3, *mult*3, *sqrt*, *power*, *negexp*, *neg*, *abs*, *log*	

4. Experimental evolution and results

4.1. Analysis of biosorption process

4.1.1. FTIR and SEM

FTIR spectra of PSH before and after TY biosorption are given in Fig. 6. The peak occurring at 3292 cm^{-1}, located between 3000–3600 cm^{-1} in the FTIR spectra, is attributed to the hydrogen bonded OH stretching vibrations (Demiral & Şamdan, 2016; Kaur, Singh, & Rajor, 2021). The peak seen at 2929 cm^{-1} represents the C–H stretch. The peak at 1743 cm^{-1} shows the stretching of the C=O carbonyl bond originating from esters and ketones (Subbaiah & Kim, 2016). The peak

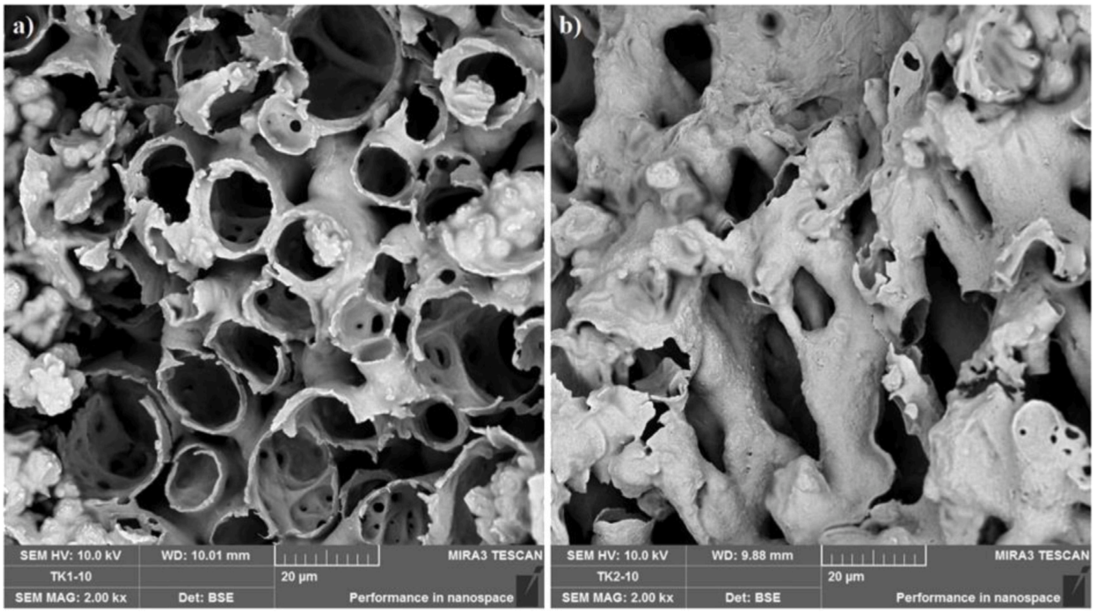


Fig. 7. PSH SEM images (a) before biosorption, (b) after biosorption.

at 1453 cm^{-1} indicates the C–H stretching and the strong peak at 1030 cm^{-1} for C–O stretching vibration, respectively (Demiral & Şamdan, 2016; Kaur et al., 2021). The strong peak that emerged in 1634 cm^{-1} refers to the C=C vibrations attributed to the olefinic and aromatic structure for the PSH. The peak occurring at 1229 cm^{-1} indicates the presence of ester groups. The peak at 557 cm^{-1} indicates the presence of alkynes and alkyl halides (Demiral & Şamdan, 2016).

SEM images of PSH before and after TY biosorption are given in Fig. 7. SEM images were obtained at 2 kx magnification and at a scale of 20 μm. The heterogeneous and porous structure of PSH was clearly visible. After the biosorption process, it was observed that the PSH surface and the pores are coated with dye (Hameed & El-Khaiary, 2008).

4.1.2. *Effect of pH on biosorption*

The pH change of the dye solution is very important for the biosorption capacity as it affects the solubility of the dye and the surface charge of the adsorbent (Subbaiah & Kim, 2016). According to Fig. 8(a), the pH_{pzc} value was determined to be 6.86. Here, the point from the graph where the initial pH value equals the final pH value is determined as pH_{pzc}. At pH conditions below this value, the biosorbent surface becomes cationic and electrostatically interacts with TY, an ionic dye. This leads to an increase in biosorption under low pH conditions (Bameri et al., 2022; Mittal, Ahmad, & Mittal, 2021; Rigueto et al., 2021). In biosorption, there is an electrostatic interaction between the biosorbent and the dye (Raju & Sunil, 2018). When examining the effect of pH value of dye solution (2–8) on biosorption, initial dye concentration was 50 mg/L and biosorbent concentration was 1 g/L with 120 min contact time (Fig. 8(b)). While the pH value of the TY solution was 2, the biosorption efficiency reached 90.7% and the biosorption capacity reached 30.49 mg/g. Percentage biosorption efficiency of 21.4%, 34%, and 19.9% at pH 4, 6 and 8, respectively, and biosorption capacity values of 9.63, 25.01 and 11.87 mg/g were achieved. A sharp decrease in efficiency and biosorption capacity is observed when pH rises from 2 to 4. The reason for this situation may be that the surface charge changes with the loss of protons on the biosorbent surface (Mittal et al., 2021). For anionic dyes, it is seen that the biosorption capacity decreases as the pH level increases. Weak electrostatic interaction might be the cause of this situation (Rattanapan et al., 2017). In addition, when the pH value is greater than 6, TY precipitates and biosorption efficiency decreases (Raju & Sunil, 2018).

4.1.3. *Effect of biosorbent concentration on biosorption*

In order to examine the effect of the concentration of biosorbent on the biosorption, it was studied at an initial dye concentration of 50 mg/L at pH 2, with the concentration of 0.5–10 g/L biosorbent. When Fig. 9 is examined, it is clearly seen that the biosorption capacity decreases with the increase in the concentration of biosorbent. Biosorption efficiency did not show any significant change with the increase of the concentration of biosorbent and biosorption efficiency values close to each other were obtained. Biosorption efficiency of 92% and 92.17%, which were close to values at 0.5 g/L and 10 g/L, was achieved, respectively. The highest biosorption capacity was achieved with 74.88 mg/g and 0.5 g/L biosorbent concentration. For this reason, 0.5 g/L biosorbent concentration was chosen and the experiments were continued. As the concentration of biosorbent increases, the number of active sites on the biosorbent surface also increases. Biosorption continues until the active sites are saturated. However, this situation causes a decrease in the mobility in the solution as the concentration increases (Azizinezhad, 2022). In addition, agglomeration occurs as the concentration of biosorbent increases (Mahmoodi et al., 2016). This result shows that high biosorption efficiency can be achieved with a low concentration of biosorbent, and that both the cost and the amount of waste will be low (Isik et al., 2022).

4.1.4. *Interpretation of isotherm models used*

The isotherm models examined in the biosorption of TY to PSH are given in Fig. 10. The R^2 value in the Langmuir isotherm was determined as 0.7124 (Fig. 10(a)). The q_{max} value was calculated as 181.8 mg/g according to the Langmuir isotherm. In addition, the R^2 value in the Freundlich isotherm was determined as 0.839 (Fig. 10(b)). The diffuse porous structure of the biosorbent surface in SEM images showed that it could have a heterogeneous surface. In the D-R isotherm (Fig. 10(c)), the R^2 value was determined as 0.9874. The constant k_D and E were calculated as 0.0047 mol^2/J^2 and 10.31 kJ/mol, respectively. In this case, it can be concluded that there is ion exchange in the biosorption process (Isik et al., 2022). The effect of initial dye concentration on biosorption efficiency and biosorption capacity is given in Fig. 10(d). The highest biosorption efficiency (%) was achieved at a dye concentration of 100 mg/L (pH:2, biosorbent concentration: 0.5 g/L, contact time: 120 min), with a biosorption capacity of 166.26 mg/g. It can be seen that with the increase of initial dye concentration, the biosorption efficiency (%) and biosorption capacity increase up to

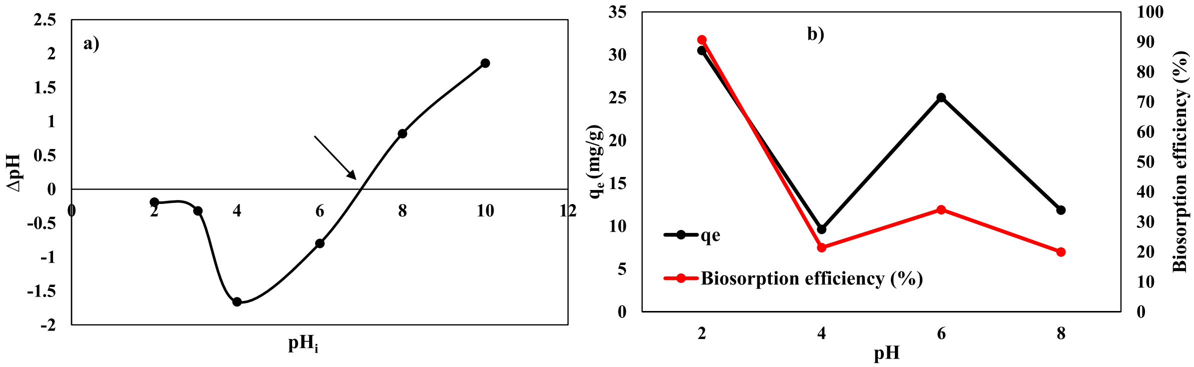


Fig. 8. Effect of pH on biosorption efficiency (%) and biosorption capacity (qe).

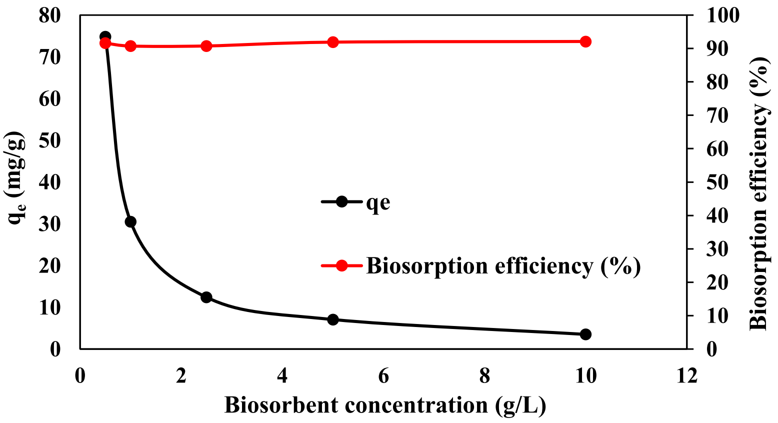


Fig. 9. Effect of biosorbent concentration on biosorption efficiency (%) and biosorption capacity (q_e).

Table 4
Various sorbents and experimental information used in the adsorption of TY in the literature.

Sorbent	Biosorption efficiency (%)	q max (mg/g)	Contact time	References
Supported chitosan adsorbent	97.95	120.48	90 min	Shi et al. (2020)
Wallnut husk	85.5	7.8431	20 min	Ibrahim et al. (2021)
Kahwa tea carbon	–	55.55	240 min	Mittal et al. (2021)
Polyaniline-SiO$_2$ nanocomposite	95	141.5	80 min	Rastgordani, Zolgharnein, and Mahdavi (2020)
Fe$_3$O$_4$	–	30	60 min	Pietrzyk et al. (2022)
Fe$_3$O$_4$@10%Zn	–	43	60 min	Pietrzyk et al. (2022)
PSH	95	181.8	120 min	This study

a certain concentration and then decrease. The reason for this situation may be insufficient active sites on the biosorbent surface with increasing concentration (Saravanan et al., 2021). Kinetic and thermodynamic studies were investigated for an initial dye concentration of 100 mg/L.

Table 4 demonstrates the utilization of agricultural wastes, composites, and nanoparticles as sorbents in TY adsorption. It can be expressed that PSH has a greater q_{max} value than synthesized composites. With this finding, it was determined that PSH, a naturally occurring, inexpensive, and plentiful agricultural waste, achieved high biosorption because of its porous structure. PSH was employed in this investigation without any chemical treatment, which saves energy and money.

4.1.5. *Kinetic models interpretation*

The relationship between contact time and dye removal efficiency in the biosorption process is known as biosorption kinetics (Shi et al., 2020). The reaction kinetics (Fig. 11) in the biosorption of TY to PSH were investigated with an initial dye concentration of 100 mg/L and a biosorbent concentration of 0.5 g/L at pH 2 for 120 min.

When Fig. 11 is examined, it is seen that the highest regression values for TY biosorption kinetics were obtained with the PSO model (R^2 = 0.9943) and the Elovich model (R^2 = 0.9524). The Elovich model constants were calculated as α 45.18 mg/g min and β 0.027 g/mg, respectively. The PSO model states that there is a physicochemical interaction between the two phases and that there is surface biosorption including chemisorption. The Elovich model, on the other hand, suggests that biosorption occurs by mass transfer and that there is chemisorption at the surface (Rigueto et al., 2021). The rate constant (k_2) for the PSO model was calculated as 0.002 g/mg.min. The numerical value of the rate constant gives information about whether the biosorption between the biosorbent and the dye reaches equilibrium quickly or slowly. The low computed k_2 value can mean that it takes the biosorption a while to reach equilibrium (Netzahuatl-Muñoz et al., 2012).

4.1.6. *Thermodynamic data interpretation*

The effect of temperature on TY dye biosorption of PSH was investigated for 20, 35 and 50 °C. The graph in Fig. 12 was obtained according to the Van't Hoff equation in Eq. (15). Eqs. (14)–(16) calculated parameters ΔH, ΔS and ΔG. The ΔH value was calculated as −30.55 kJ/kmol and the ΔS value as −79.32 J/mol K. According to the calculated ΔH value, biosorption is a physical and exothermic process (Isik et al., 2022). The ΔG value calculated for 20, 35 and 50 °C temperatures was

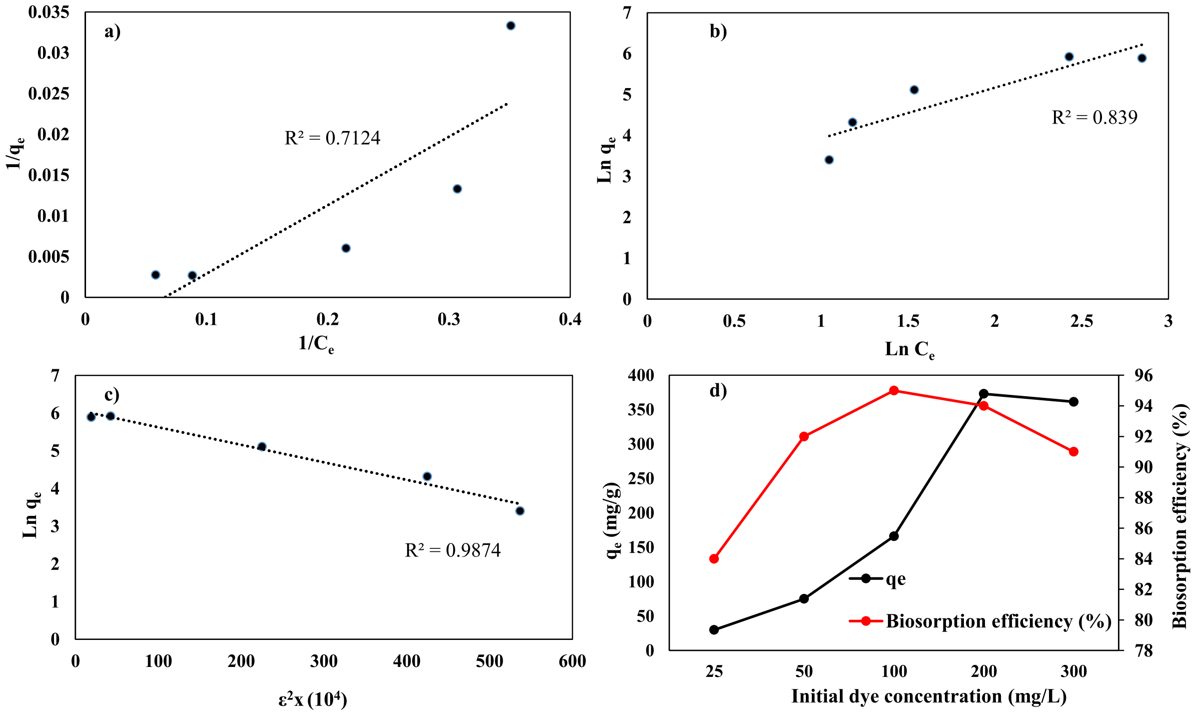


Fig. 10. (a) Langmuir, (b) Freundlich, (c) D-R Isotherms. (d) Effect of initial dye concentration on biosorption efficiency (%) and biosorption capacity (q_e).

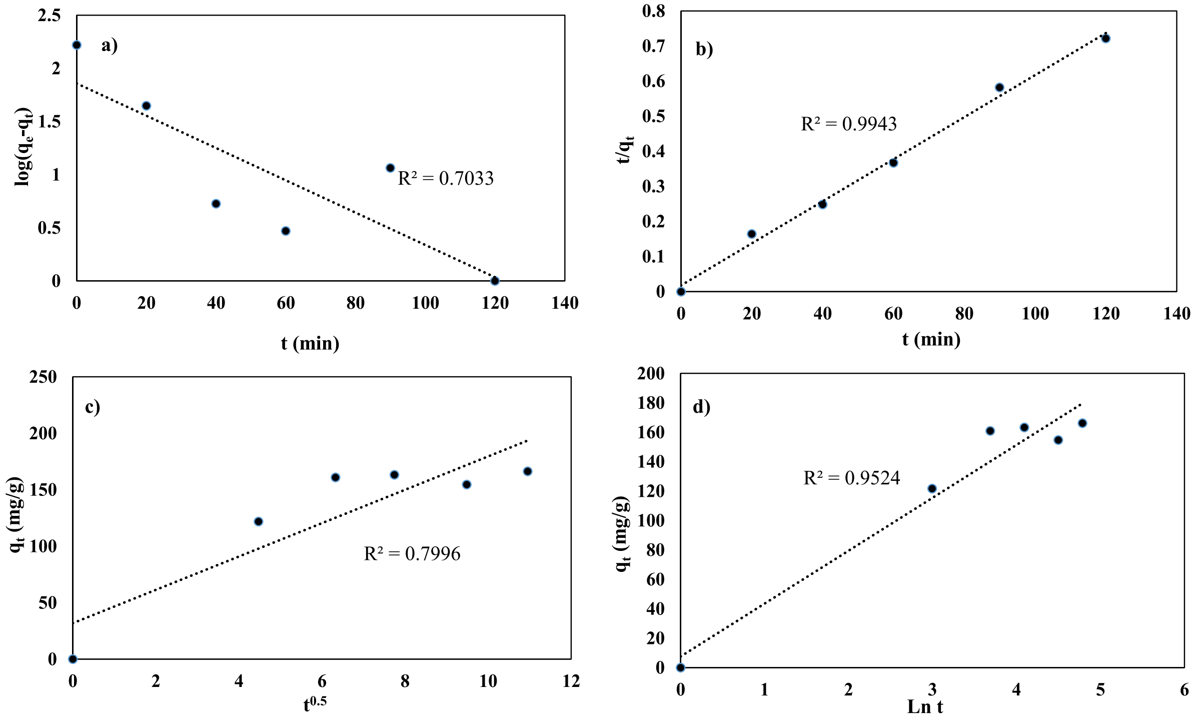


Fig. 11. Biosorption kinetics, (a) PFO model, (b) PSO model, (c) Intraparticle diffusion, (d) Elovich model.

determined as −7.3, −6.1 and −4.91 kJ/mol, respectively. A negative ΔG value indicates that biosorption occurs spontaneously (Joudi et al., 2020).

Following the evaluation of the collected data, several inferences regarding the biosorption process between PSH and TY were made. According to theory, the PSH surface protonates in an acidic environment, and the anionic dye TY then initiates the biosorption process. We can discuss the existence of chemical adsorption because the exothermic reaction is compatible with PSO. Additionally, the biosorption's compliance to the D-R isotherm suggests that ion exchange has occurred. In Fig. 13, the biosorption experiment is given schematically. The change in color of PSH before and after biosorption clearly indicates that the biosorbent has biosorbed the dye. In addition, the color of the TY solution became transparent after biosorption. After this part of the study, by modeling the biosorption process with EC-based AP methods, the parameters affecting the biosorption were investigated with empirical equations.

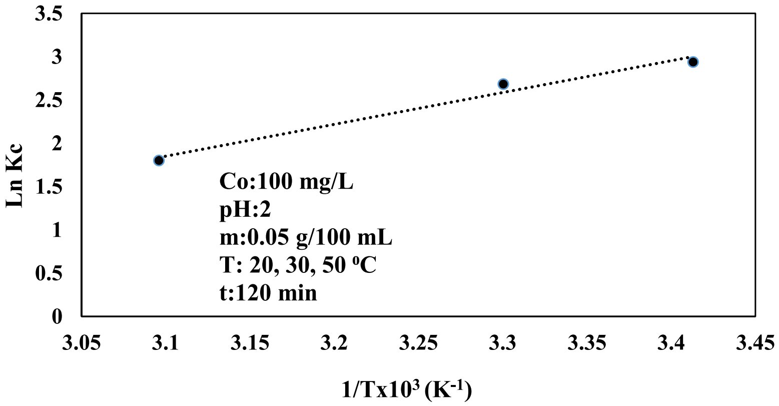


Fig. 12. Effect of temperature on PSH biosorption of TY (Van’t Hoff graph).

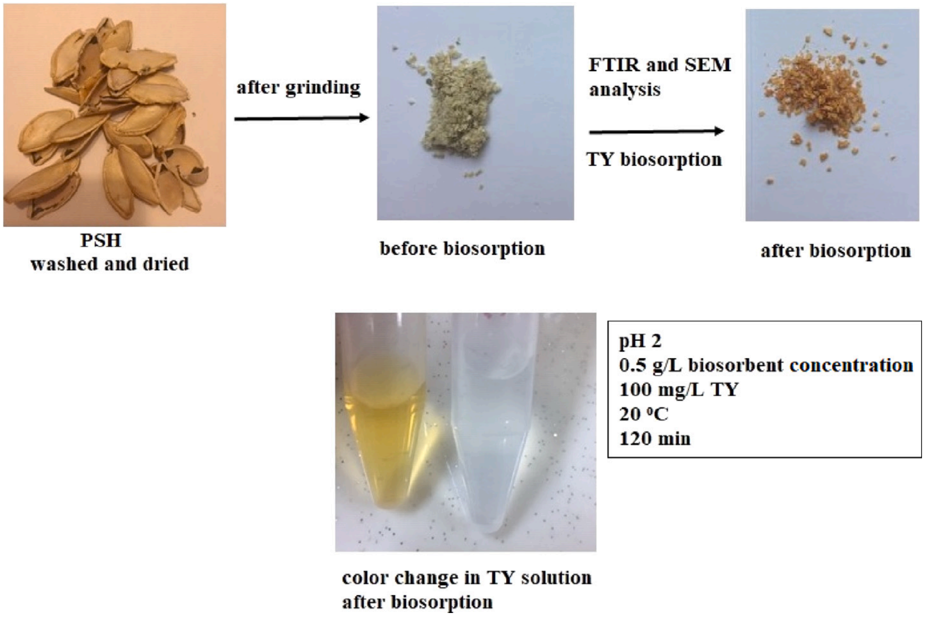


Fig. 13. Schematic view of the biosorption process.

4.2. Simulation results

4.2.1. Results of the AP methods

The modeling studies were tested with 100 independent runs for each method and the results were shown in Table 5. The dataset resulting from the experiments was first randomly mixed to learn the methods using different samples. The table contains the results of the models generated by both methods for both training and test data. For $RMSE_{\text{Train}}$, MAE_{Train} and R^2_{Train}, the models from both methods obtained very similar results. The significant difference in the results of the methods was seen in the test data. The mean value of MGGP models for SSE_{Test} is approximately 1.14 times higher than that of MHABCP. However, this difference is not observed in the R^2 values of the models. When evaluating the best models, the model of MHABCP is quite low compared to the MGGP model both in terms of number of nodes and complexity. What is striking in the results in this table is that the MHABCP models have lower error values and higher R^2 values than MGGP for all evaluation criteria.

Because the performance evaluation criteria and R^2 values of the two methods were very close, one-sided t-tests were performed to determine whether there was a significant difference between the methods. The null hypothesis that the mean of MGGP’s performance evaluation criteria was equal to that of MHABCP was tested with the alternative that the mean was greater than that of MHABCP. The significance level (α) used for the test was set at 0.05. The results of the t-tests for all criteria present in Table 6.

It can be noted from the Table 6 that the returned value of $h = 0$ accepts the null hypothesis that the means of MGGP and MHABACP are equal. This means that there is no significant difference at the 5% level for the performance criteria tested. From this table, it can be seen that for the criteria SSE_{Train}, SSE_{Test}, MAE_{Train}, there is a significant difference between the two methods and the null hypothesis is rejected. In addition, the null hypothesis was not accepted for the total number of nodes and complexity. The t-test results in Table 6 support the simulation results in Table 5. It is worth noting that there is a significant difference in favor of MHABCP for some performance criteria.

4.2.2. Analysis of the best models

Table 7 shows the comparative results of the best models with the lowest error and the highest R^2_{Test} values obtained by both methods after 100 runs.

In general, the MHABCP model has lower errors than the MGGP model. However, the R^2_{test} value of MGGP is slightly higher than that of MHABCP. Both models have very high R^2 values in both training and testing. The most significant difference between the criteria is seen in

Table 5
Simulation results.

Method	MHABCP	MGGP	MHABCP	MGGP	MHABCP	MGGP	MHABCP	MGGP
Criteria	Mean		Best		Worst		Standart deviation	
$RMSE_{\text{Train}}$	2.3076	2.6062	2.0188	2.3893	3.8312	3.9672	0.4754	0.5712
$RMSE_{\text{Test}}$	10.3412	11.1470	6.1626	5.8459	15.8888	17.2983	2.6439	2.5112
SSE_{Train}	327.4954	356.0730	80.1573	85.5713	866.0008	928.5871	137.8742	169.3837
SSE_{Test}	2848.2636	3262.4791	854.3779	949.4488	6311.3789	7480.7909	1451.2575	1436.5562
MAE_{Train}	1.4968	1.6575	0.9316	0.8347	2.6180	2.7636	0.3276	0.4324
MAE_{Test}	6.8905	7.4148	4.4464	4.3823	11.0314	10.2585	1.5293	1.4113
R^2_{Train}	0.9965	0.9962	0.9992	0.9991	0.9908	0.9902	0.0015	0.0018
R^2_{Test}	0.9558	0.9176	0.9784	0.9760	0.9239	0.8110	0.0123	0.0361
$Number of Nodes$	71.7000	80.2000	42.0000	55.0000	110.0000	123.0000	15.5345	13.9068
$Complexity$	198.1400	234.1700	96.0000	151.0000	360.0000	384.0000	53.5785	47.9268

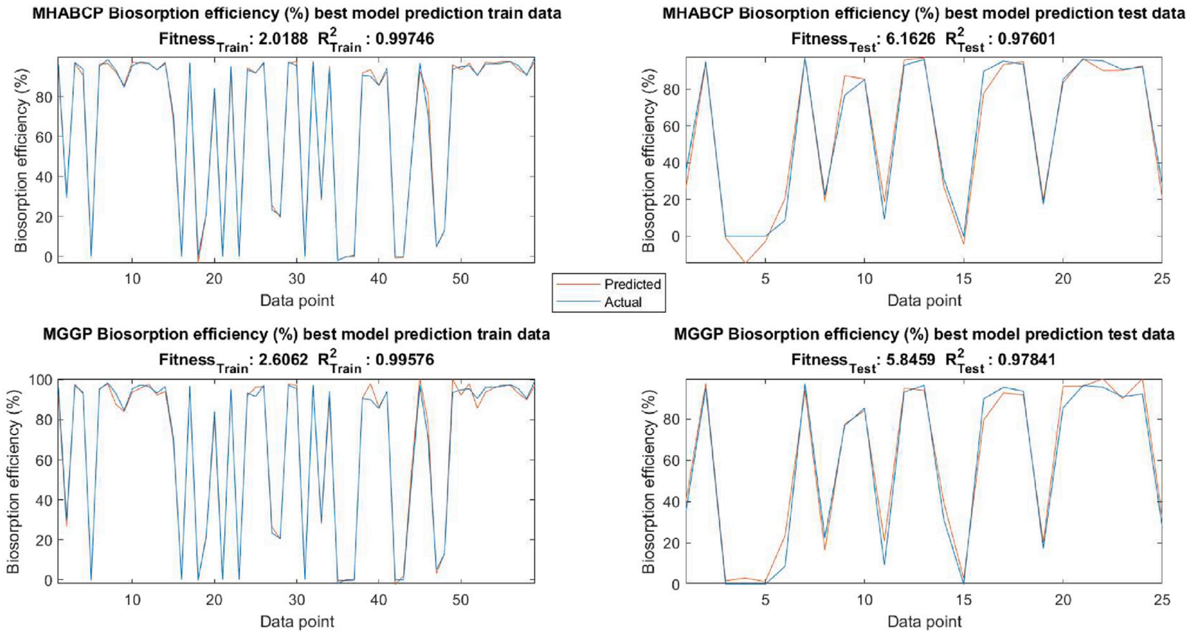


Fig. 14. The relationship between the predicted and actual data.

Table 6
h and p-value of t-tests.

	h	p
$RMSE_{\text{Train}}$	0	0.1296
$RMSE_{\text{Test}}$	0	0.1722
SSE_{Train}	1	0.0010
SSE_{Test}	1	3.8E–05
MAE_{Train}	1	0.0014
MAE_{Test}	0	0.1592
R^2_{train}	0	0.9099
R^2_{test}	0	1.0000
$Number of Nodes$	1	5.6E–06
$Complexity$	1	3E–07

the SSE_{Test} values. According to these values, MGGP is approximately 1.7 times higher than MHABCP. At the same time, the model created by MGGP is about 0.5 times more complex than the MHABCP model. This shows that there is no direct relationship between the complexity of the model and the accuracy of the prediction. Simplified formulas of the best models in Table 7 are shown in Table 8.

The dependent/independent parameters in Table 8 are presented in Section 3.1. Since the complexity of the MGGP model is greater than that of the MHABCP. Therefore, it can be said that it is more advantageous to prefer MHABCP. Moreover, both formulas contain all dependent parameters. In addition, Figs. 14 and 15 are plotted to see how well they fit the biosorption data while analyzing the best models in this study.

Fig. 14 shows how the methods relate to training and testing data. The figure shows that data from chemical experiments can be modeled effectively using EC -based AP methods. The models fit the training data better than the test data. However, both models have a R^2_{test} value of about 0.98. As the R^2 values decrease by 2% between training and test data, the predictive accuracy of the models also decreases. Both methods fit the training and test data well with high R^2 values, and there is no overfitting (memorization of the training data and failure of the test data). This is also confirmed by the scatter plots in Fig. 15.

Fig. 15 is a scatter plot where the value of the actual values determines the position on the horizontal axis and the value of the predicted values determines the position on the vertical axis. The closer the blue points on the scatter plots are to the lines, the stronger the correlation between the actual and predicted values. In this figure, the points converged with the line when training the best MHABCP model. The convergence is greater for the training data than for the test data. Considering Figs. 14 and 15 together, MGGP and MHABCP are a great success.

4.2.3. *Feature analysis results in the biosorption process*

In this subsection, feature selection with SR was performed to understand which features are most relevant/necessary to the biosorption process. For this purpose, the frequencies of the input parameters were calculated and visualized in the distributions in all models with high R^2 values of 100 runs in both methods. The frequency distributions for the models with $R^2 \geq 0.8$ and $R^2 \geq 0.9$ are given in Fig. 16. The population

Table 7
Comparative results of the best models.

Method	$RMSE_{\text{Train}}$	$RMSE_{\text{Test}}$	SSE_{Train}	SSE_{Test}	MAE_{Train}	MAE_{Test}	R^2_{Train}	R^2_{Test}	Best model criteria	
MHABCP	2.0188	6.1626	80.1573	240.4629	1.2928	4.4464	0.9975	0.9760	Nodes Depth Complexity	69 4 205
MGGP	2.3893	5.8459	85.5713	400.7300	1.8136	4.3823	0.9958	0.9784	Nodes Depth Complexity	110 4 334

Table 8
Simplified formulas.

Method	Simplified formulas
MHABCP	$y = 1554.0\,\mathrm{e}^{-1.0\,\mathrm{e}^{-1.0\,x_1} - 1.0\,\cos(x_4) - \frac{1.0\,x_4}{x_1}} - 117.6\,\mathrm{e}^{-\frac{1.0\,{x_4}^2}{x_3+2.0\,x_4}}$ $-8549.0\,\mathrm{e}^{-\frac{1.0\,x_3\,x_4}{2.0\,x_4+x_5}} - 47.84\,\mathrm{e}^{-1.0\sqrt{\frac{x_4}{x_3}}}$ $+1527.0\,\mathrm{e}^{-1.0\,\mathrm{e}^{-\frac{1.0\,x_4}{x_1}}} + 9052.0\,\mathrm{e}^{-\frac{1.0\,{x_4}^2}{{x_1}^2}} - 4.173\,\cos\left(x_3\,\cos\left(x_5\right)\,\left(x_1 + x_5 - 6.176\right)\right)$ $-3915.0\,\mathrm{e}^{-1.0\,\mathrm{e}^{-1.0\,{x_1}^2}} - \frac{2.742e{+}15\,\cos(x_1)}{1.407e{+}14\,x_3 - 1.407e{+}14\,x_4} - 3.415\sqrt{\frac{x_4\,x_5}{x_3+7.405}} + 2443.0$
MGGP	$y = 4.816\,\cos\left(\sqrt{x_1 + x_2 + x_4}\right) - 0.3403\,x_3 - 0.3403\,\sin\left(x_1 + x_3 + x_4\right) - 0.3403\,\sin\left(2.0\,x_2 + x_5\right)$ $-4.163\,\cos\left(\frac{x_3}{\cos(x_4)}\right) - 13.67\,x_1 - 58.01\,\cos\left(x_3 - 0.7854\right)$ $+27.76\,\sin\left(x_1\right) - 0.6806\,\sin\left(x_5\right) - 693.2\,\mathrm{e}^{-1.0\,x_1 - 2.0\,x_4} + \frac{6.939e{-}18\left(4.197e{+}15\,x_3 + 8.394e{+}15\,\sin(x_5)\right)}{\cos({x_4}^2)}$ $-\frac{2.22e{-}16\left(5.26e{+}15\,x_1 + 2.63e{+}15\,\sin(x_5)\right)}{\cos(x_4)}$ $+0.008989\,x_5\,\cos\left(x_1\right)\,\left(2.0\,x_1 + x_4 + \cos\left(x_4\right) + \sin\left(x_5\right)\right) + 0.007106\,x_1\,\cos\left(x_4\right)\,\left(x_2 + x_4 + x_5\right)$ $+0.002369\,\cos\left(x_4\right)\,\left(2.0\,x_2 + x_4\right)\,\left(x_1 + x_2 + x_5\right) + 148.4$

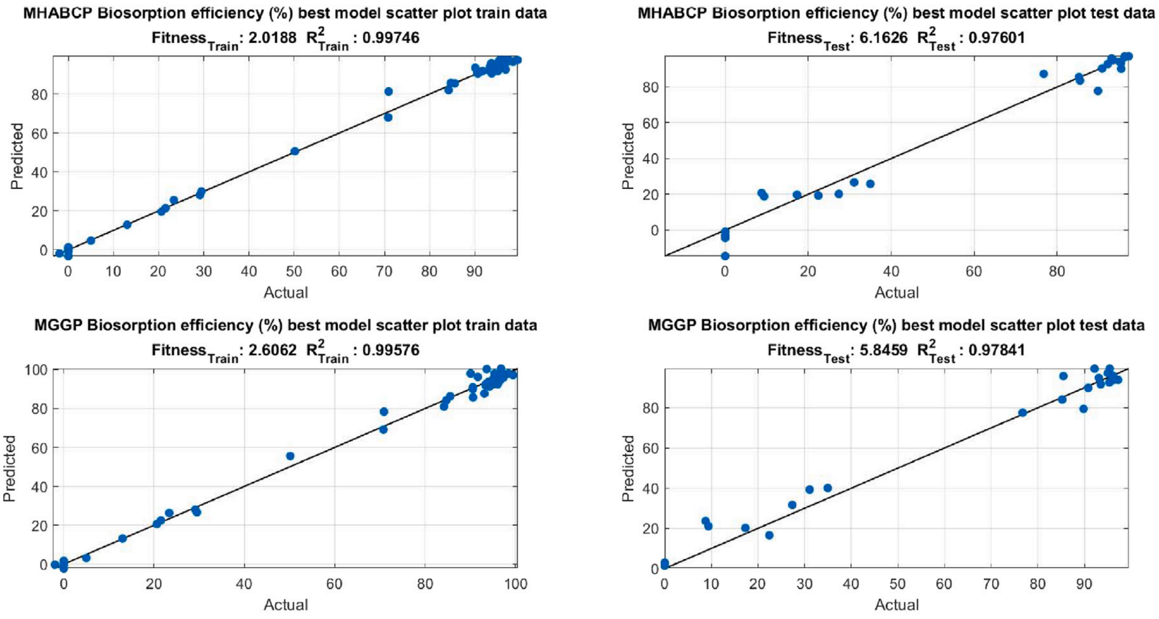


Fig. 15. The scatter plots between the predicted and actual data.

size/colony size of each run is 500. Therefore, the population size of a total of 100 runs is 500 * 100 = 50 000. In other words, Fig. 16 was obtained by analyzing all runs and evaluating a total of 50 000 different individuals.

As mentioned in Section 3.1, the input parameters are (x_1) pH, (x_2) biosorbent concentration (g/L), (x_3) initial dye concentration (mg/L), (x_4) contact time (min), (x_5) temperature (°C), and (y) biosorption efficiency (%). Fig. 16 clearly shows that the methods contain the same inputs in the formulas at similar frequency frequencies. The most significant point that MHABCP has more individuals with higher R^2 values for all population members. Another important point is how significant the inputs are, from high to low, respectively, x_4, x_1, x_3, x_5, x_2.

4.2.4. Additional study comparing AP methods

Our goal is to compare the success of the AP method with another method. For this purpose, we have chosen the newly proposed Interaction-Transformation Evolutionary Algorithm (ITEA) for SR problems as our benchmarking algorithm. There are three important reasons why we chose this algorithm:

- The algorithm is based on the interaction transformation representation, unlike the AP methods.
- It has a closed form for the gradient of the generated model, thanks to its differential representation, which allows to compute the partial effect on the returned expressions.

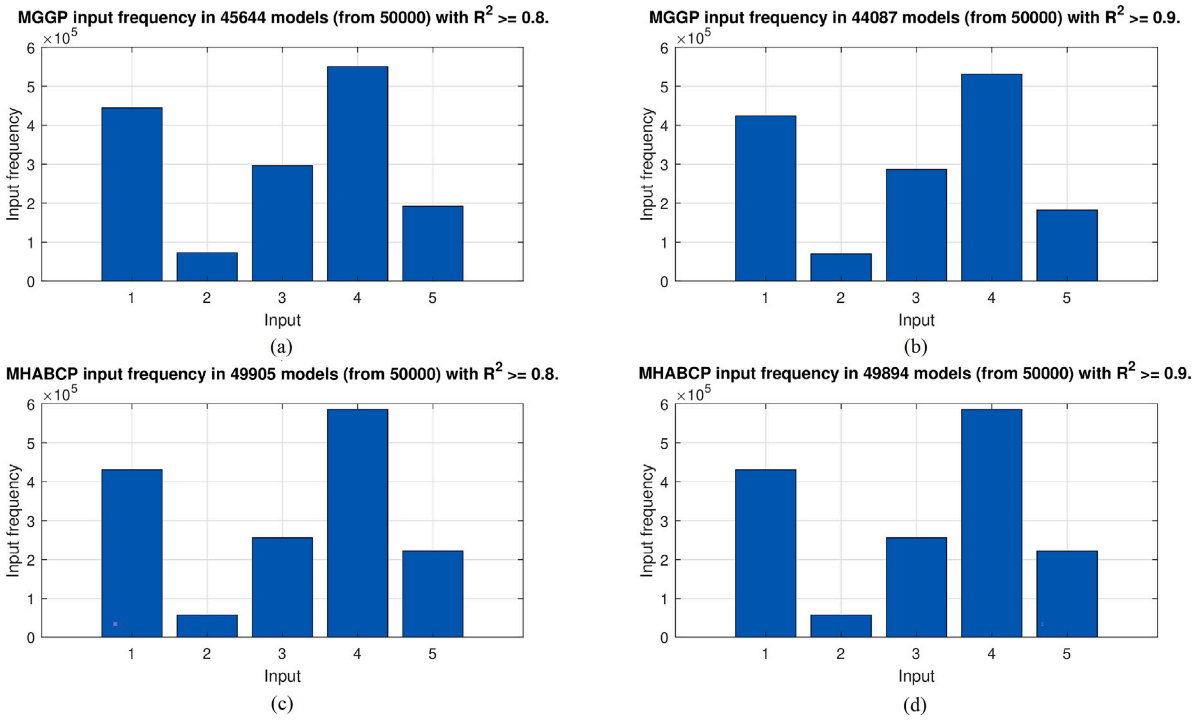


Fig. 16. Frequency distribution of input variables.

Table 9
Comparison of the AP methods and ITEA.

Method	MHABCP		MGGP		ITEA	
	$RMSE_{Train}$	$RMSE_{Test}$	$RMSE_{Train}$	$RMSE_{Test}$	$RMSE_{Train}$	$RMSE_{Test}$
Mean	2.3076	10.3412	2.3893	11.1470	3.3006	12.7008
Best	2.0188	6.1626	2.6062	5.8459	2.8701	6.5911
Worst	3.8312	15.8888	3.9672	17.2983	5.1357	38.1795
Standard deviation	0.4754	2.6439	0.5712	2.5112	0.7942	6.9792

Table 10
Best model of ITEA.

$RMSE_{Train}$	$RMSE_{Test}$	Simplified model
2.8701	6.5911	$0.023\sqrt{x_1 x_2^3 x_3^2 x_4^2 x_5^3} - 307.895\sqrt{x_2^3 x_4 x_5} - 54.508\tan\left(x_2^{-3} x_3 x_4^{-2}\ \ x_5^{-3}\right)$ $- 0.046\sqrt{x_1^{-1} x_2^3 x_3^2 x_4^2 x_5^3} + 25.404e^{x_1^3 x_2^2 x_3 x_4^{-2}} + 54.483$ $+ 883.37\sqrt{x_1^{-3} x_2^3 x_4 x_5} + 16.452e^{x_1^3 x_2^{-2} x_3 x_4^{-3} x_5^{-1}}$

- ITEA was able to generate better models than the algorithms random forest (RF), extreme gradient boosting (XGB) and kernel ridge regression (KR) (Aldeia & de França, 2021).

To make a fair comparison, the parameter values given in Table 3 were not changed in ITEA. Therefore, the population size used in the experiments is 250 and the number of iterations is 500. The maxterms 10, strength range $(-3, 3)$ is taken from the specific parameters of the algorithm. The results of 100 independent runs are shown in Table 9.

The most striking aspect of this table is that ITEA gives results that are very close to the AP methods. The best mean and $RMSE_{Train}$ is that of MHABCP. ITEA is about 1 higher than the mean and about 0.8 higher than the best $RMSE_{Train}$. The biggest difference between the methods of ITEA and AP methods is the height of the standard deviation in the test data. This shows that the difference between the best and worst model produced by the algorithm is significant. The comparison results prove that ITEA is a strong competitor for the AP methods. The best model produced by this algorithm for the data is given in Table 10. The model is produced using square root and exponent functions rather than trigonometric functions.

5. Discussion

The answers to the research questions based on our experiments and simulations are listed below.

- Can PSH be used as an efficient biosorbent for the removal of TY dye from wastewater?
  Answer: It is thought that PSH will be used as an efficient biosorbent due to the results it achieves in a short contact time.
- Under which experimental conditions can the highest biosorption efficiency be achieved when the process is optimized?
  Answer: For the biosorption of TY dye from wastewater by PSH, 95% biosorption efficiency was achieved at pH 2, 0.5 g/L biosorbent concentration, 100 mg/L initial dye concentration and 120 min contact time.
- Which isotherm and kinetic models is the biosorption reaction compatible with?
  Answer: The biosorption process follows the D-R isotherm and the PSO kinetic model. The reaction occurs spontaneously.

- Can EC-based AP methods be used to model chemical experiments such as biosorption?
  Answer: These methods fit the data well, with very high R^2 values for both training and test data. Moreover, the high success of the models in the test is evidence that the methods do not overfitting the data. Considering the process as a SR problem and producing its mathematical models means that we can identify the process without additional experimental costs. Therefore, this study shows that EC-based AP methods can be used in modeling biotechnological applications.
- What is the observed error value when the biosorption process is mathematically modeled using different methods?
  Answer: In the study, we assessed the models based on three different performance criteria. We also calculated the number of nodes and the complexity of the models to analyze the trees, and the R^2 values to evaluate the fit to the data. According to the simulation results, the SSE values are the highest among the criteria and the R^2 values are close to 1.
- In what order are the features that contribute to the process important?
  Answer: To evaluate the features, the models with the highest R^2 values were analyzed. Below is the order of importance of the features for the biosorption process according to the analysis, from high to low:
  - contact time
  - pH
  - initial dye concentration
  - temperature
  - biosorbent concentration

6. Conclusions

In this study, the biosorption of TY was investigated by optimizing the process using unmodified PSH, which is a cheap, easily available, and natural sorbent. In addition, mathematical models were generated by using the data of the biosorption process obtained from experimental studies using MGGP and MHABCP. The success of the methods is influenced by the values of the initial parameters. Therefore, in our study, the optimal parameters for the methods had to be tuned. For this purpose, we tried to find the most optimal parameters among 516 different combinations by performing hyperparameter optimization. In addition, a biosorption efficiency of 95% was achieved in the parameter ranges selected for the biosorption process, which was optimized for pH 2–8, 0.5–10 g/L biosorbent concentration, 10–300 mg/L initial dye concentration, and 0–120 min contact time.

The findings are as follows:

- At pH 2, with an initial dye concentration of 100 mg/L, and a biosorbent concentration of 0.5 g/L, 95% biosorption efficiency was achieved after 120 min.
- A q_{max} value of 181.8 mg/g was calculated according to the Langmuir isotherm.
- According to D-R isotherm, it is found that PSH has a heterogeneous surface and is subject to multilayer biosorption.
- Biosorption of TY to PSH is an exothermic reaction and occurs spontaneously.
- According to the simulation results, it should be noted that MAHBCP generates better models compared to MGGP.
- The success and complexity of the models are analyzed, and the effects of the parameters on the process are examined.
- We compared AP methods and the newly proposed ITEA in an additional study. The results show that these methods are better than ITEA.
- According to the models, the most significant parameter is the contact time.

When all the findings are evaluated together, it can be concluded that EC-based AP methods can be used in modeling and analysis of real problems such as the biosorption process. We plan to evaluate the success of using EC -based methods in modeling various chemical processes in future studies. We also plan to improve the performance of hybrid methods and select features for high-dimensional problems.

CRediT authorship contribution statement

Sibel Arslan: Methodology, Software, Validation, Writing – reviewing, Supervision. **Nurşah Kütük:** Methodology, Experimental, Investigation, Data analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

References

Adeyi, O., Adeyi, A. J., Oke, E. O., Okolo, B. I., Olalere, A. O., Otolorin, J. A., et al. (2022). Process integration for food colorant production from Hibiscus sabdariffa calyx: A case of multi-gene genetic programming (MGGP) model and techno-economics. *Alexandria Engineering Journal*, *61*(7), 5235–5252, URL: https://www.sciencedirect.com/science/article/pii/S1110016821006931.

Aldeia, G. S. I., & de França, F. O. (2021). Measuring feature importance of symbolic regression models using partial effects. In *Proceedings of the genetic and evolutionary computation conference* (pp. 750–758).

Aliwi, M., Demirci, S., & Aslan, S. (2023). Difference-based firefly programming for symbolic regression problems. *Computer Standards & Interfaces*, Article 103722.

Alizadeh, M. J., Shahheydari, H., Kavianpour, M. R., Shamloo, H., & Barati, R. (2017). Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network. *Environmental Earth Sciences*, *76*(2), 1–11.

Arslan, S., & Koca, K. (2023). Investigating the best automatic programming method in predicting the aerodynamic characteristics of wind turbine blade. *Engineering Applications of Artificial Intelligence*, *123*, Article 106210, URL: https://www.sciencedirect.com/science/article/pii/S0952197623003949.

Arslan, S., & Ozturk, C. (2018). Artificial bee colony programming for feature selected cancer data classification. *International Journal of Scientific & Technology Research*, *4*(7), 75–84.

Arslan, S., & Ozturk, C. (2019a). Artificial bee colony programming descriptor for multi-class texture classification. *Applied Sciences*, *9*(9), URL: https://www.mdpi.com/2076-3417/9/9/1930.

Arslan, S., & Ozturk, C. (2019b). Multi hive artificial bee colony programming for high dimensional symbolic regression with feature selection. *Applied Soft Computing*, *78*, 515–527.

Azizinezhad, F. (2022). Surface adsorption of Pb^{2+}ions from aqueous solutions using chitosan grafted with a mixture of IA-MAm/bentonite. *International Journal of Environmental Analytical Chemistry*, 1–17.

Bahramian, M., Dereli, R. K., Zhao, W., Giberti, M., & Casey, E. (2022). Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Systems with Applications*, Article 119453.

Bameri, I., Saffari, J., Baniyaghoob, S., & Ekrami-Kakhki, M.-S. (2022). Synthesis of magnetic nano-NiFe$_2$O$_4$ with the assistance of ultrasound and its application for photocatalytic degradation of Titan Yellow: Kinetic and isotherm studies. *Colloid and Interface Science Communications*, *48*, Article 100610.

Boudouaoui, Y., Habbi, H., Ozturk, C., & Karaboga, D. (2020). Solving differential equations with artificial bee colony programming. *Soft Computing*, *24*(23), 17991–18007.

Chaurasia, P., Jasuja, N. D., & Kumar, S. (2022). A sustainable approach in bioremediation of textile dye effluent by microbial consortia. *International Journal for Research in Applied Science and Engineering Technology*, *10*, 868–876.

Datta, S., Dev, V. A., & Eden, M. R. (2019). Developing non-linear rate constant QSPR using decision trees and multi-gene genetic programming. *Computers & Chemical Engineering*, *127*, 150–157, URL: https://www.sciencedirect.com/science/article/pii/S0098135419300602.

Demiral, İ., & Şamdan, C. A. (2016). Preparation and characterisation of activated carbon from pumpkin seed shell using H$_3$PO$_4$. *Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering*, *17*(1), 125–138.

Ge, J., Yusa, N., & Fan, M. (2021). Frequency component mixing of pulsed or multi-frequency eddy current testing for nonferromagnetic plate thickness measurement using a multi-gene genetic programming algorithm. *NDT & E International*, *120*, Article 102423, URL: https://www.sciencedirect.com/science/article/pii/S0963869521000220.

Guo, Z., Hu, S., Han, Z.-K., & Ouyang, R. (2022). Improving symbolic regression for predicting materials properties with iterative variable selection. *Journal of Chemical Theory and Computation*, *18*(8), 4945–4951.

Hadi, S. J., & Tombul, M. (2018). Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. *Journal of Hydrology*, *561*, 674–687, URL: https://www.sciencedirect.com/science/article/pii/S0022169418302890.

Hameed, B., & El-Khaiary, M. (2008). Removal of basic dye from aqueous medium using a novel agricultural waste material: Pumpkin seed hull. *Journal of Hazardous Materials*, *155*(3), 601–609.

Hiremath, S., Mal, A. R., Prabha, C., & Vidya, C. (2018). Tamarindus indica mediated biosynthesis of nano TiO_2 and its application in photocatalytic degradation of titan yellow. *Journal of Environmental Chemical Engineering*, *6*(6), 7338–7346.

Ibrahim, H. K., Allah, M. A. A. H., & Muneer, A. (2021). Adsorption of titan yellow using walnut husks: Thermodynamics, kinetics and isotherm studies. *Annals of the Romanian Society for Cell Biology*, *25*(6), 12576–12587.

Isik, B., Ugraskan, V., & Cankurtaran, O. (2022). Effective biosorption of methylene blue dye from aqueous solution using wild macrofungus (Lactarius piperatus). *Separation Science and Technology*, *57*(6), 854–871.

Joudi, M., Nasserlah, H., Hafdi, H., Mouldar, J., Hatimi, B., El Mhammedi, M., et al. (2020). Synthesis of an efficient hydroxyapatite–chitosan–montmorillonite thin film for the adsorption of anionic and cationic dyes: Adsorption isotherm, kinetic and thermodynamic study. *SN Applied Sciences*, *2*(6), 1–13.

Karaboga, D., Ozturk, C., Karaboga, N., & Gorkemli, B. (2012). Artificial bee colony programming for symbolic regression. *Information Sciences*, *209*, 1–15.

Kaur, G., Singh, N., & Rajor, A. (2021). Adsorption of doxycycline hydrochloride onto powdered activated carbon synthesized from pumpkin seed shell by microwave-assisted pyrolysis. *Environmental Technology and Innovation*, *23*, Article 101601.

Kazemi, M., & Barati, R. (2022). Application of dimensional analysis and multi-gene genetic programming to predict the performance of tunnel boring machines. *Applied Soft Computing*, *124*, Article 108997, URL: https://www.sciencedirect.com/science/article/pii/S1568494622003222.

Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, *4*(2), 87–112.

Kütük, N., & Arslan, S. (2022). Biosorption of methyl orange from aqueous solution with hemp waste, investigation of isotherm, kinetic and thermodynamic studies and modeling using multigene genetic programming. *Chemical Papers*, *76*(12), 7357–7372.

Mahmoodi, N. M., Chamani, H., & Kariminia, H.-R. (2016). Functionalized copper oxide–zinc oxide nanocomposite: synthesis and genetic programming model of dye adsorption. *Desalination and Water Treatment*, *57*(40), 18755–18769.

May Tzuc, O., Hernández-Pérez, I., Macias-Melo, E., Bassam, A., Xamán, J., & Cruz, B. (2019). Multi-gene genetic programming for predicting the heat gain of flat naturally ventilated roof using data from outdoor environmental monitoring. *Measurement*, *138*, 106–117, URL: https://www.sciencedirect.com/science/article/pii/S0263224119301502.

Mehdizadeh, S., Mohammadi, B., & Ahmadi, F. (2022). Establishing coupled models for estimating daily dew point temperature using nature-inspired optimization algorithms. *Hydrology*, *9*(1), 9.

Mirjalili, S. (2016). Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications*, *27*, 1053–1073.

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software*, *95*, 51–67.

Mittal, J., Ahmad, P., & Mittal, A. (2021). Kahwa tea (Camellia sinensis) carbon—a novel and green low-cost adsorbent for the sequestration of titan yellow dye from its aqueous solutions. *Desalination Water Treat*, *227*, 404–411.

Moazenzadeh, R., Mohammadi, B., Duan, Z., & Delghandi, M. (2022). Improving generalisation capability of artificial intelligence-based solar radiation estimator models using a bio-inspired optimisation algorithm and multi-model approach. *Environmental Science and Pollution Research*, *29*(19), 27719–27737.

Moghaddam, S. A. V., Al-Sahaf, H., Xue, B., Hollitt, C., & Zhang, M. (2021). An automatic feature construction method for salient object detection: A genetic programming approach. *Expert Systems with Applications*, *186*, Article 115726.

Mohammadi, B. (2023). Modeling various drought time scales via a merged artificial neural network with a firefly algorithm. *Hydrology*, *10*(3), 58.

Moreira, V. R., Lebron, Y. A. R., & de Souza Santos, L. V. (2020). Predicting the biosorption capacity of copper by dried chlorella pyrenoidosa through response surface methodology and artificial neural network models. *Chemical Engineering Journal Advances*, *4*, Article 100041.

Nekoei, M., Moghaddas, S. A., Golafshani, E. M., & Gandomi, A. H. (2021). Introduction of ABCEP as an automatic programming method. *Information Sciences*, *545*, 575–594.

Netzahuatl-Muñoz, A. R., Guillén-Jiménez, F. d. M., Chávez-Gómez, B., Villegas-Garrido, T. L., & Cristiani-Urbina, E. (2012). Kinetic study of the effect of pH on hexavalent and trivalent chromium removal from aqueous solution by Cupressus lusitanica bark. *Water, Air and Soil Pollution*, *223*(2), 625–641.

Oh, S. (2022). Predictive case-based feature importance and interaction. *Information Sciences*, *593*, 155–176.

Öztürk, C., Tarım, M., & Arslan, S. (2020). Feature selection and classification of metabolomics data using artificial bee colony programming (ABCP). *International Journal of Data Mining and Bioinformatics*, *23*(2), 101–118.

Pawanr, S., Garg, G. K., & Routroy, S. (2022). Prediction of energy consumption of machine tools using multi-gene genetic programming. *Materials Today: Proceedings*, *58*, 135–139, URL: https://www.sciencedirect.com/science/article/pii/S2214785322001833. International Conference on Artificial Intelligence & Energy Systems.

Pedrino, E. C., Yamada, T., Lunardi, T. R., & de Melo Vieira, J. C. (2019). Islanding detection of distributed generation by using multi-gene genetic programming based classifier. *Applied Soft Computing*, *74*, 206–215, URL: https://www.sciencedirect.com/science/article/pii/S1568494618305738.

Petersen, B. K., Landajuela, M., Mundhenk, T. N., Santiago, C. P., Kim, S. K., & Kim, J. T. (2019). Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. arXiv preprint arXiv:1912.04871.

Pietrzyk, P., Phuong, N. T., Olusegun, S. J., Hong Nam, N., Thanh, D. T. M., Giersig, M., et al. (2022). Titan yellow and congo red removal with superparamagnetic iron-oxide-based nanoparticles doped with zinc. *Magnetochemistry*, *8*(8), 91.

Punugupati, G., Kandi, K. K., Bose, P., & Rao, C. (2017). Process modeling of gelcast Si3N4ceramics using multi gene genetic programming. *Materials Today: Proceedings*, *4*(2), 1900–1909.

Raju, C., & Sunil, K. (2018). Studies on biosorption of titan yellow dye with hypnea musciformis powder and optimization through central composite design. *International Journal of Innovative Science and Research Technology*, *5*(1), 2349–6010.

Rastgordani, M., Zolgharnein, J., & Mahdavi, V. (2020). Derivative spectrophotometry and multivariate optimization for simultaneous removal of titan yellow and bromophenol blue dyes using polyaniline@ SiO_2 nanocomposite. *Microchemical Journal*, *155*, Article 104717.

Rattanapan, S., Srikram, J., & Kongsune, P. (2017). Adsorption of methyl orange on coffee grounds activated carbon. *Energy Procedia*, *138*, 949–954.

Rengasamy, D., Mase, J. M., Kumar, A., Rothwell, B., Torres, M. T., Alexander, M. R., et al. (2022). Feature importance in machine learning models: A fuzzy information fusion approach. *Neurocomputing*, *511*, 163–174.

Rengasamy, D., Rothwell, B. C., & Figueredo, G. P. (2021). Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Applied Sciences*, *11*(24), 11854.

Rigueto, C. V. T., Alessandretti, I., da Silva, D. H., Rosseto, M., Loss, R. A., & Geraldi, C. A. Q. (2021). Agroindustrial wastes of banana pseudo-stem as adsorbent of textile dye: characterization, kinetic, and equilibrium studies. *Chemistry Africa*, *4*(4), 1069–1078.

Rivero, D., Fernandez-Blanco, E., & Pazos, A. (2022). DoME: A deterministic technique for equation development and symbolic regression. *Expert Systems with Applications*, *198*, Article 116712.

Saravanan, A., Karishma, S., Kumar, P. S., Varjani, S., Yaashikaa, P., Jeevanantham, S., et al. (2021). Simultaneous removal of Cu (II) and reactive green 6 dye from wastewater using immobilized mixed fungal biomass and its recovery. *Chemosphere*, *271*, Article 129519.

Saravanan, A., Kumar, P. S., Yaashikaa, P. R., Kanmani, S., Varthine, R. H., Muthu, C. M. M., et al. (2019). Modelling on the removal of dye from industrial wastewater using surface improved enteromorpha intestinalis. *International Journal of Environmental Research*, *13*(2), 349–366.

Sattar, M., Majid, A., Kausar, N., Bilal, M., & Kashif, M. (2022). Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences. *Computational Biology and Chemistry*, *98*, Article 107638, URL: https://www.sciencedirect.com/science/article/pii/S1476927122000184.

Shi, Q.-X., Li, Y., Wang, L., Wang, J., & Cao, Y.-L. (2020). Preparation of supported chitosan adsorbent with high adsorption capacity for titan yellow removal. *International Journal of Biological Macromolecules*, *152*, 449–455.

Subbaiah, M. V., & Kim, D.-S. (2016). Adsorption of methyl orange from aqueous solution by aminated pumpkin seed powder: Kinetics, isotherms, and thermodynamic studies. *Ecotoxicology and Environmental Safety*, *128*, 109–117.

Tabaraki, R., & Sadeghinejad, N. (2018). Comparison of magnetic Fe_3O_4/chitosan and arginine-modified magnetic Fe_3O_4/chitosan nanoparticles in simultaneous multidye removal: experimental design and multicomponent analysis. *International Journal of Biological Macromolecules*, *120*, 2313–2323.

Teodorovic, D., Lucic, P., Markovic, G., & Dell'Orco, M. (2006). Bee colony optimization: principles and applications. In *2006 8th seminar on neural network applications in electrical engineering* (pp. 151–156). IEEE.

Vidya, C., Manjunatha, C., Sudeep, M., Ashoka, S., & Raj, L. A. (2020). Photo-assisted mineralisation of titan yellow dye using ZnO nanorods synthesised via environmental benign route. *SN Applied Sciences*, *2*(4), 1–15.

Wei, G., Zhao, J., Feng, Y., He, A., & Yu, J. (2020). A novel hybrid feature selection method based on dynamic feature importance. *Applied Soft Computing*, *93*, Article 106337.

Yamashita, G. H., Fogliatto, F. S., Anzanello, M. J., & Tortorella, G. L. (2022). Customized prediction of attendance to soccer matches based on symbolic regression and genetic programming. *Expert Systems with Applications*, *187*, Article 115912.

Yang, X.-S. (2008). *Nature-inspired metaheuristic algorithms*. Luniver Press.

Yang, X.-S., & Deb, S. (2009). Cuckoo search via Lévy flights. In *2009 world congress on nature & biologically inspired computing (NaBIC)* (pp. 210–214). IEEE.

Zhang, H., Zhou, A., Chen, Q., Xue, B., & Zhang, M. (2023). SR-forest: A genetic programming based heterogeneous ensemble learning method. *IEEE Transactions on Evolutionary Computation*.

Zojaji, Z., Ebadzadeh, M. M., & Nasiri, H. (2022). Semantic schema based genetic programming for symbolic regression. *Applied Soft Computing*, *122*, Article 108825.